UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

CREACIÓN, IMPLEMENTACIÓN Y EVALUACIÓN DE UN MODELO DE DIFUSIÓN DE INFORMACIÓN EN UNA RED SOCIAL

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

MIGUEL EMILIO GUTIÉRREZ ÁLVAREZ

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
PABLO ROMÁN ASENJO
FELIPE AGUILERA VALENZUELA

SANTIAGO DE CHILE
OCTUBRE 2012

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR : MIGUEL EMILIO GUTIÉRREZ ÁLVAREZ
FECHA: 29/10/12
PROF. GUIA: PROF. SEBASTIÁN RÍOS PÉREZ

CREACIÓN, IMPLEMENTACIÓN Y EVALUACIÓN DE UN MODELO DE DIFUSIÓN DE INFORMACIÓN EN UNA RED SOCIAL

El auge de Internet se enmarca dentro del progreso continuo de los medios de comunicación masiva, siendo el sucesor de innovaciones anteriores como la radio y la televisión. En particular, el subfenómeno de las redes sociales virtuales ha cobrado una enorme importancia desde hace ya casi una década. Millones de personas se inscriben e interactúan a diario en sitios web como facebook y Twitter — para citar a las más conocidas — compartiendo emociones, opiniones, noticias, fotografías, entre otros.

Ahora bien, desde la perspectiva científica, la comprensión del problema de la difusión de información en redes sociales virtuales es aún parcial, debido a lo reciente de su inicio. Por otro lado, en relación a su atractivo comercial, se ha difundido la idea que las redes sociales pueden ser una nuevo canal de ventas relevante para el futuro, incluso indispensable. Aún así, las empresas están incursionando tímidamente en ellas, no siempre con éxito.

Luego, modelos explicativos acerca de la dinámica de las redes sociales pueden ser de gran ayuda, para incentivar la inversión de las empresas en estrategias efectivas de marketing en redes sociales. En ese contexto, el objetivo principal de este trabajo consiste en crear, implementar y desarrollar un modelo de difusión de información en redes sociales.

Para lograr lo anterior, un modelo de interacción de usuarios en un foro web fue creado e implementado, bajo la forma de un simulador en tiempo continuo. Este simulador emula las decisiones que toman los usuarios mediante la incorporación de un modelo de decisión perceptual proveniente de las neurociencias. Datos reales de un foro web chileno, `http://www.plexilandia.cl/foro`, fueron empleados para la calibración requerida por el simulador, entre Noviembre de 2009 y Marzo de 2010. Luego, cinco meses de actividad fueron simulados, entre Abril y Agosto de 2010, y comparados con los datos reales.

Se analizaron aspectos como la generación de contenidos, la generación de grafos y la difusión de información. Las métricas de desempeño utilizadas fueron el error porcentual absoluto promedio (MAPE), la medida F (F-measure) y la pendiente de la estimación por mínimos cuadrados ($\hat{b}$), respectivamente. Además del modelo principal antes mencionado, se incorporaron modelos adicionales a modo de benchmark.

El modelo principal obtuvo en el óptimo un MAPE semanal promedio de 7.321%, un F-measure de 2.687% y una pendiente de $7.32 \times 10^{-7}\, semana^{-1}$. A modo de trabajo futuro se sugiere perfeccionar la estructura modelada del foro y explorar variaciones adicionales en los parámetros del modelo.

Para Valeska.

Agradecimientos

Agradezco a mi Profesor Guía, Dr. Sebastián Ríos, por darme la oportunidad de realizar este trabajo, y a mi Profesor Co-Guía, Dr. Pablo Román, por la guía brindada en la implementación de los modelos. Este trabajo fue posible gracias a una base de datos y web logs del sitio `http://www.plexilandia.cl/foro` que fueron brindados por el Dr.(c) Felipe Aguilera, y pertenece al proyecto Fondecyt de iniciación, código 11090188, llamado "Semantic Web Mining Techniques to Study Enhancements of Virtual Communities".

Gracias a mis compañeros de oficina por el tiempo que hemos compartido. Gracias a Luciano Villarroel por la energía que infunde al equipo, a Daniel Beth, Julio Quinteros y Carlos Reveco por su valiosa ayuda en computación, a Ricardo Muñoz por los papers e información útil que ha compartido, y a Iván Videla por su sentido del humor.

Muchas gracias a Lautaro Cuadra por haberme recomendado, gracias a él pude realizar este trabajo. Gracias a mis amigos: la familia Cariz, la familia Valdivia, Adrián Albala, Joel Olmos, Hernán Castro, Mauricio Durán, Gabriel Bravo, Fernando Badilla y Gustavo Pavez.

Gracias a mis primos y primas, por todo lo que hemos pasado juntos. Un abrazo para Graham Blanc, Pablo Valenzuela, Felipe Martínez, Amaranta Martínez, Sergio Munizaga, Daniela Munizaga, Constanza Barra, Martina Barra, Amelia Barra, Diego Álvarez, Pascal Álvarez, Camila Álvarez, Rolando Gutiérrez y Javiera Gutiérrez.

Un abrazo para mis tías y tíos Claudia Gutiérrez, Domingo Barra, Alfonso Gutiérrez, Rebeca Caceres, Maria Inés Fuentealba, Leticia Álvarez, Sergio Valenzuela, María Isabel Álvarez, Sergio Munizaga, Edmundo Álvarez y Blanca Garretón. Esta memoria también está para recordar a mis abuelos Lola, Rolando, Edmundo e Yvette, y a Guillermo quién nos dejó este año.

Un gran abrazo para mi mamá Roxana, mi papá Guido y mi hermano Felipe. Este trabajo es el fruto de años de cariño, de acción y de reacción, y de amistad y compañerismo. Los quiero mucho, sin ustedes no lo habría logrado.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Since almost ten years, online social networks (OSN) have brought a revolution in mass media communications. Not only they are being increasingly adopted in a wide scale around the world and across age segments, but they also enable a two-way interaction to an extent just unseen before. As a result, an increasing number of activities in the virtual world are becoming *social*: social media, social business, social customer relationships management, and so on. As a concrete example, social media played an important role in the 2011-2012 Egyptian Revolution [5]. In Chile, a well-known retailer was the first company to reach one million "likes" on facebook, in August 2012 [4]. Recently, facebook reached one billion users worldwide on 2012/10/04 [3]. OSNs, and in general social media, are definitely a hot topic nowadays.

Yet, social networks have existed for a very long time. From our ancestral tribes to modern world great metropolis, human societies still exhibit a natural tendency toward self organization in complex social structures. These structures reflect a simple fact. On one hand, not all members have the same role. On the other hand, members of a social system do not interact with each other with the same intensity, and some preferential relationships appear. On the academic side, sociology — by means of social network analysis (SNA) — has been studying social networks since the beginnings of the 20th century second half, long before the rise of Internet. SNA is a sociological research method that merges sociological theory with graph theory. Indeed, graphs are used as its primary tool to model interaction and relationships in human networks.

The OSN data has allowed a significant increase in studies sample sizes. For example, the famous experiment of Travers and Milgram on the small-world phenomenon was conducted on a set of only 296 selected individuals [43]. Nowadays, with the advent of OSN (of which the most prominent examples are facebook and Twitter), there are now massive amounts of new available evidence. The latter reveal with fine-grained level of detail the structure of human networks around the globe (figure 1.1). For example, the 2005 Leskovec et al. study of an online recommendation network [32] considered a 4 million people sample size. Even better, the 2012 Backstrom et al. study of the distribution of degrees of separation in the

facebook graph [14] was conducted with information of 721 million users.

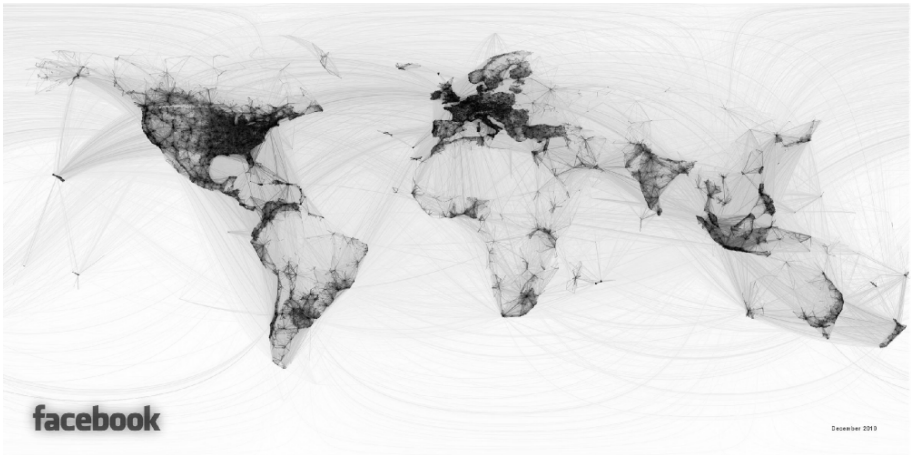


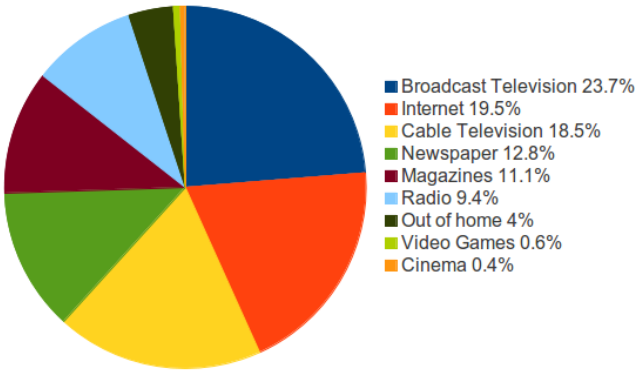
Figure 1.1: View of the friendships network in facebook, as of 2010. Each pair of cities are connected by a great circle arc, which depends on their euclidean distance and the number of mutual friendships involved. *Source: facebook engineering team note [7].*

The social structure plays a very important role in the social system to which it belongs. It is closely interrelated with the distribution of power among the members, with their statuses and their roles. In particular, in the case of the diffusion of an innovation among a social system, very different results will be obtained if the adoption-promoting effort is directed either to well-connected members or to peripheral and marginal ones [37, p. 2]. Intuitively, the best results would be expected in the former case: each member can spread innovation to a greater number of community's members, with greater effectiveness. Modern applications, such as viral marketing for example, are laid upon this kind of intuition [19, p. 21].

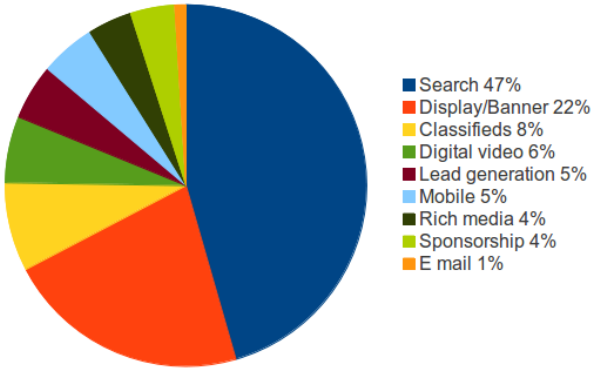
The novelty of this work is the use of a model recently developed in the field of perceptual choice (the leaky competing accumulator model), as the underlying mechanism in a diffusion process applied to text documents, modeled as vectors obtained from latent semantic algorithms (such as latent Dirichlet allocation LDA). In this introductory chapter, some basic principles on the information diffusion through social networks are exposed. Then research hypotheses, work's objectives and expected results are raised, in order to define the formal framework. Finally, report structure is presented.

1.1 Background

In this section, facts about the online marketing industry are provided, some conceptual foundations are laid down and useful related work is reviewed. The goal is to stress the importance of the research problem, and to sketch some possible paths leading to a solution.



(a) Advertising revenue market share by media - 2011



(b) Online advertising revenue market share by format - 2011

Figure 1.2: Market shares of global and online advertising in the US for year 2011. *Source: IAB internet advertising revenue report [1].*

1.1.1 The online marketing industry

In order to have a point of reference, the total advertising revenues are analysed first for the United States of America. During year 2011, the global advertising market generated revenues equal to 162.3 billion USD [1]. The Internet represents 19.5% of the global advertising revenues, with 31.7 billion USD. It constitutes the second most important media after broadcast television [1]. From all media considered in the analysis, only two exhibit a positive compound annual growth rate (CAGR) during period 2005–2012: Internet with 16.7%, and cable television with 4.0% [1]. Therefore, Internet is actually the fastest growing media in terms of annual revenue, ahead of other communication channels. Regarding the composition of Internet revenues, the two most important items are search engine ads (47%) and display/banners (22%) [1]. Nonetheless, careful attention should be paid to the mobile revenues. Indeed, although they rank low with 1.6% billion USD, it is the fastest growing item with a 149% growth over 2010–2011 [1].

Yet, the online marketing practice in social media still suffers from drawbacks. There is a perceived absence of meaningful metrics about return on investment (ROI) [21, 23]. Moreover, "companies invest millions of dollars in social media, with little understanding of how it influences consumers to favor their brands or buy their products" [23]. The result is

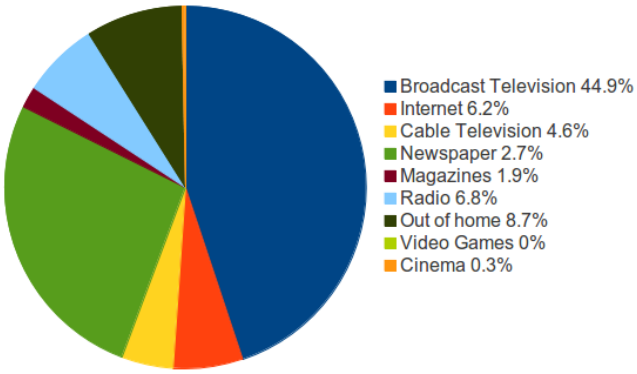
that social media efforts still account for 1% of an average marketing budget, although their potential is being increasingly recognized [23]. Nonetheless, a clearer conceptual framework is emerging gradually, and four principal functions of social media may be identified: to monitor, to respond, to amplify and to lead consumer behavior [23]. Moreover, social media are the only form of marketing that influence the consumer at every stage of the decision process, from brand/product awareness to actual purchase [23]. If now social media are the third most used online channel after company home page and e-mail, it is expected to become second in one to three years, behind mobile applications and ahead of company home page [21].

Concerning the Chilean market, as of 2011 there is a total of $2,025,066$ fixed connections to the Internet, which represents a households penetration of 38%, as well as a population penetration of 58% [2]. There are $7,957,714$ mobile connections (46% of the population), of which 38% are 3G connections [2]. While the number of fixed connections has grown 11.3% from 2010 to 2011, the number of mobile connections has more than duplicated with a 104.8% growth [2]. 93% of Chileans with access to the Internet make use of online social networks, the two following uses being search (89%) and multimedia with 77% [2]. In particular, facebook has a penetration of 90.7%, Twitter of 13.8% and the average Chilean Internet user spends 8.8 hours monthly in online social networks [2].

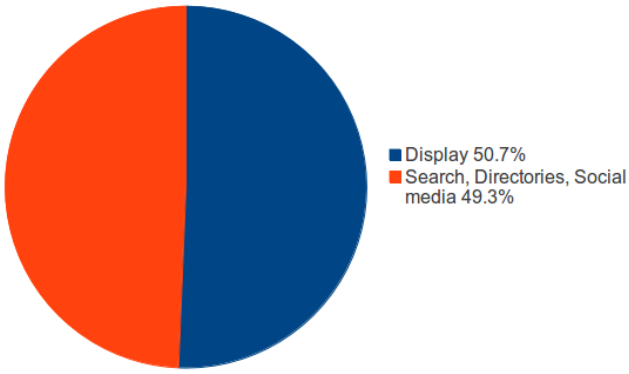
During year 2011, 1.36 billion USD have been spent in advertising, where 85 million USD (6.2%) went to online advertising [2]. It is the media with the third fastest growing spending share, after cable television and out of home. From 2010 to 2011, online advertising revenues grew by 30% [2]. However, the online advertising spending share is lesser than in other countries with similar or even lower penetration [2]. Although Internet is the media where users spend most time in relative terms, it is also the media with one of the lowest global advertising revenues share [2]. Therefore, it can be expected that, similarly as in the US case, the online advertising will grow on a sound basis in then next years, and a great deal of attention should be paid to both social network and mobile. However, in the author's experience, the lack of metrics is a significant problem for online marketers too in Chile. There is therefore a potential market for meaningful social networks metrics.

1.1.2 The information diffusion through social networks problem

Online marketing practitioners invest on social networks advertising with the goal of generating revenues. Comparing the advertising costs with the resulting incomes, the ROI is computed: the higher the ROI, the better. ROI is obviously important, but true knowledge of consumers could be more. On the other hand, the current practice of online social networks monitoring is focused on counting. The temporal analysis of online *buzz* is based on time series. Both the ROI measurement and the buzz analysis lay on a "black box" assumption: the relation between advertising budget and generated incomes, and the temporal evolution of the volume of mentions may be known, but not the underlying mechanism. Therefore, It would be useful for online marketing practitioners to support their decisions with a standard



(a) Advertising revenue market share by media - 2011



(b) Online advertising revenue market share by format - 2011

Figure 1.3: Market shares of global and online advertising in Chile for year 2011. *Source: Internet en Chile [2].*

model of information spread through online social networks (OSN). Moreover, it could be a source of competitive advantage in the near future.

The research problem concerns the mechanism of such black box, and can be explicitly stated as follows:

How does a social network behave when exposed to new information?

Indeed, the knowledge of how a social network reacts to new information could lead to the elaboration of better advertising contents, more suited to companies' marketing goals. The research problem is sketched in figure 1.4. In the previous statement, three key concepts are mentioned: social network, behavior and information. These are defined below.

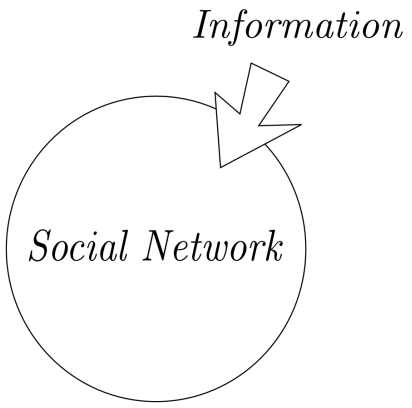


Figure 1.4: Sketch of the research problem: given a social network, what happens if some new information is introduced into it?

The concept of social network is related to the emergence of the *social network perspective* [45, p. 4] in the social sciences, in particular during the second half of 20^{th} century. The social network perspective is rooted in the critique of prior social science methods, which assumed that individuals were mostly independent, with no consideration of interaction patterns. This was particularly evident in the case of the diffusion of innovations, and as noted by Rogers, "the individual has been largely used as the unit of analysis in diffusion research, rather than the sociometric dyad, network, clique, unit more appropriate for investigating the *process* aspects of diffusion" [37, p. 80]. A few decades after Rogers, Wasserman and Faust retrieved the following principles:

- "Actors and their actions are viewed as interdependent rather than independent, autonomous units
- Relational ties (linkages) between actors are channels for transfer or "flow" of resources (either material or non material)
- Network models focusing on individuals view the network structural environment as providing opportunities for or constraints on individual action

- Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relation among actors" [45, p. 4]

More precisely, concerning the definition of a social network:

> "The relational structure of a group or larger social system consists of the pattern of relationships among the collection of actors. The concept of a network emphasizes the fact that each individual has ties to other individuals, each of whom in turn is tied to a few, some, or many others, and so on. The phrase "social network" refers to the set of actors and the ties among them". [45, p. 9]

Therefore, the inclusion of relational concepts and metrics is central in the understanding of social networks. It follows that a social network can be mathematically modeled as a graph. A graph $G = (V, E)$ is a collection of nodes V (the "set of actors"), linked by a set of edges $E = \{(u, v) : u, v \in V\}$ (the "ties").

Behavior refers to evolution over time of a set of well-defined attributes of the system under study. As, in this case, the system is a social network, then straightforward attributes are the set of nodes V and of edges E themselves. For example, one may wonder about the effects of the inoculation of a new topic of conversation during an informal meeting of a group of friends. An edge is drawn — for example — between two friends if they have talked to each other for more than 20 minutes. This introduced topic may be controversial, causing some friends to leave (V changes), or, in contrast, a nice topic that fosters conversation and invite the friends to stay, thus densifying the set of edges (E changes). Moreover, attributes may be assigned to both nodes and edges. A node attribute in this case may be a sentiment indicator, and an edge attribute may be its communication capacity. Then, the controversial topic may cause a negative shift of sentiment indicators and a decrease in the communication capacity of edges, while quite the contrary should be observed for the case of a nice topic.

Information is conceived here as the formal aspect of the broader phenomenon of communication [22, p. 15–16]. The sense, value, truth, objective, irony, and so on, of a given message are therefore not considered [22, p. 16]. This is only partially true for the sense of a message, as text documents are treated in this work as vector representations in a semantic space. Nonetheless, the emphasis is placed on the numeric aspects of these vectors, and once conversion from text documents to vectors is made, little consideration is given to the actual sense of the vector. Information may be thought of as specificity. For example, suppose someone needs to find a book in a library with N books, where n are blue [22, p. 55]. The higher the ratio N/n is, the higher the informational value of the sentence "the book is blue" is [22, p. 55]. If one half of books is blue, uncertainty is reduced by one half; if one tenth is blue, it is reduced by nine tenths [22, p. 55]. In the context of this work, a higher component of a topic component will denote a higher degree of belonging of a text document to the corresponding topic.

A social network is thus seen as a system which undergoes an evolution as the time passes. Without being exposed to new information, it would follow its *natural* evolution,

which depends on the sole inner characteristics of the system. But when the contrary happens, probably the input of information will affect the course of events, and in this case the system follows a *perturbed* evolution. Therefore, another way of stating the research problem posed before is (see figure 1.5): in which ways do the natural and the perturbed evolution of a social network differ?

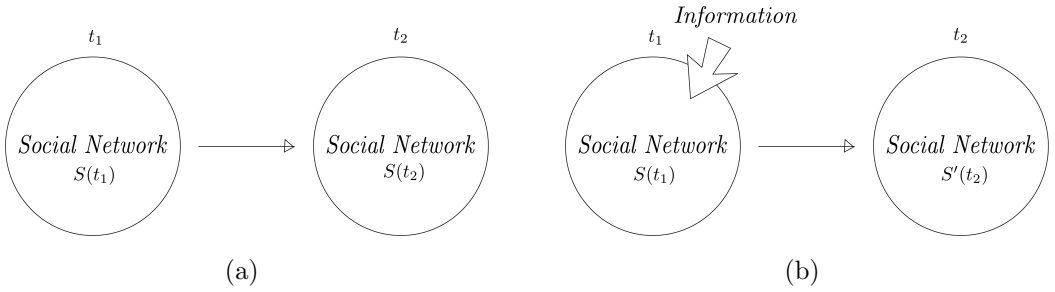


Figure 1.5: In the case (a), the social network evolves from $S(t_1)$ to $S(t_2)$ on its own. But in the case (b), it evolves from $S(t_1)$ (the same initial state than before) to $S'(t_2)$. How different are then $S(t_2)$ and $S'(t_2)$?

In order to answer this question, the attributes under consideration of an online social network must be defined. These are the set of nodes, the set of edges, the state of each node and the state of each edge. It is important to note that these attributes must be measurable, so they can be collected from actual data, analysed and even simulated. The evolution of the social network is four-fold: the nodes may change, the edges may change, the nodes' states may change, and the edges' states may change (figure 1.6).

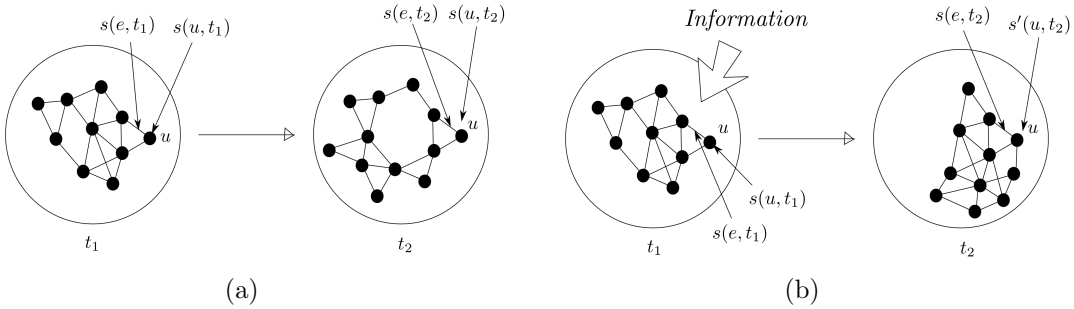


Figure 1.6: The state of the social network $S(t)$ at t is now characterized by its set of nodes, its set of edges, the state of each node, and the state of each edge. In the case (a), each node u is initially at state $s(u,t_1)$, and then naturally evolves to state $s(u,t_2)$. But in the case (b), some nodes appear or disappear (the same applies to the edges), and the node u evolves to the perturbed state $s'(u,t_2)$, as a consequence of the input of information. The same applies to each edge e.

Scope of analysis

Only the natural evolution of a OSN is studied in this work, leaving the case of perturbed evolution for future work. The primary aspect of concern is the change of the users' states, followed by the change in the edges between users. The users' states refer to the information (posts) produced during a period t. No attributes are assigned to the edges, and in addition, no variation in the users set is considered. The scope of analysis is summarized in table 1.1.

	Changes in:			
Evolution type	Nodes set V	Nodes attributes	Edges set E	Edges attributes
natural	No	Yes	Yes	No
perturbed	No	No	No	No

Table 1.1: Scope of analysis of the present work, signaling the aspects encompassed.

1.1.3 Related work

This work is an attempt of merging various research disciplines, such as the diffusion of innovations, epidemiology, information retrieval and perceptual choice, with the goal of creating a coherent hybrid algorithm that may be useful in the future. Some aspects of these traditions are summarized in the following sections.

Diffusion of Innovations

The diffusion of innovations deals with the spread of new ideas or products within a social system. Although the natural evolution of a social system (see scope of analysis, subsection 1.1.2) considers no innovation by definition, this tradition of research provides some useful concepts. Three classic publications of the field are mentioned below, ranging from 1890 to 1969.

For the purposes of this report, the starting point — concerning the tradition of research on diffusion of information in societies — is the book by french sociologist Gabriel Tarde called *The Laws of Imitation* [41] [1]. Tarde states that imitation is the defining element of human societies, upon which social ties emerge, and may be either positive (*imitation*) or negative (*counter-imitation*). He therefore refutes the utilitarian vision that affirms that a society "is a group of distinct individuals who render one another mutual services" [41, p. 59], and on the contrary argues : "Social relations, I repeat, are much closer between individuals who resemble each other in occupation and education, even if they are competitors, than between those who stand most in need of each other" [41, p. 64]. Citing some scientific evidence available at the time, he places the roots of human imitation in the human brain, "a *repeating* organ for the senses" which is "itself made up of elements which repeat on another"

[1] Originally published in 1890 as *Les Lois de l'Imitation* [42].

[41, p. 74].

In this context, each imitative behavior (a custom, an idea, a technology, etc.) can be seen as a propagating wave which progresses through the social environment [41, p. 22,49]. Concerning the *origin* of a propagating imitation (or consolidated, or even fallen into disuse), Tarde concludes that there must exist a focal point it can be traced back to, thus introducing the concept of *invention*. On the other hand, regarding its *destination*, he remarks that "when wants or ideas are once started, they always tend to continue to spread of themselves in a true geometric progression" [41, p. 115]. Moreover, he anticipates the famous S-shaped curves usually found in diffusion studies, characterized by a slow initial progress followed by a quick increment and a final slowdown [41, p. 127].

The next cornerstone is the work of Everett M. Rogers, *Communications of Innovations: A Cross-Cultural Approach* [37] [2]. Rogers systematizes much of the concepts developed by Tarde and builds up a theoretical framework for the study of the diffusion of innovations, summarizing some of the main findings of diffusion research at the time (in particular in rural and in medical sociology). The main theme developed in his book is that "*communication is essential for social change*" [37], which occurs following three distinct sequential steps: invention, diffusion and consequences [37, p. 6–7]. Rogers defines them as follows: "*invention* is the process by which new ideas are created or developed", "*diffusion* is the process by which these new ideas are communicated to the members of a social system" and "*consequences* are the changes that occur within a social system as a result of the adoption or rejection of the innovation" [37, p. 7]; so he concludes that "social change is therefore an effect of communication" [37, p. 7].

Among other aspects, Rogers focuses on the characteristics of innovations [37, p. 22] and on the innovation-decision process [37, p. 23]. Concerning the former, Rogers notes that the success of an innovation, measured by its *rate of adoption*, depends on its characteristics, of which the most important are: relative advantage, compatibility with existing values and beliefs, complexity of understanding and use, trialability, observability [37, p. 23]. Regarding the innovation-decision process, Rogers distinguishes four steps: "(1) knowledge, (2) persuasion, (3) decision, and (4) confirmation" [37, p. 25]. Therefore, combining the characteristics of an innovation with the characteristics of the members of the *social system* under study, insights may be found about why do the rates of adoption of different innovations among different people differ. The separation of knowledge and decision in the process explains the sometimes large time lags observed between awareness of a product and its adoption [37, p. 16–17]. Furthermore, Rogers calls *innovativeness* this time lag at the individual level, and classifies adopters into his five well-known categories: "(1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards" [37, p. 27].

In 1969, seven years after the publication of the first edition of Rogers' work, the paper *A New Product Growth Model For Consumer Durables* by the academic Frank M. Bass appeared in Management Science [16]. Following the tendency towards a more formal analysis of the diffusion of innovations, Bass proposes a mathematical model focused on the timing

[2] Originally published in 1962 as *Diffusion of Innovations* [36].

of the adoption of a new product (and therefore on sales through time) in a market. He distinguishes two categories of adopters, *innovators* and *imitators*, inspired by the work of Rogers: the former corresponds to the first category, while the latter is the aggregation of categories (2) to (5) [16, p. 216]. Then, the basic assumption of the model is posed, which states that "the probability that an initial purchase will be made at T given that no purchase has yet been made is a linear function of the number of previous buyers" [16, p. 216]. The previous probability depends on three parameters, *i.e.* a coefficient of innovation, a coefficient of imitation and the size of the market. Taking that assumption for granted, Bass then derives mathematically a continuous model of product adoption, as well as a discrete analogue, and performs a regression on the data provided by eleven products in the U.S.A.

Epidemiology

Another form of diffusion is the spread of an epidemic among a susceptible population. The field of epidemiology can be of no use in sociological terms, but the mathematical formulation is indeed of great help for more general diffusion phenomena, as acknowledged by Bass for example [16, p. 215]. In the following, the publications that marked the birth of modern epidemiology are presented.

Between 1927 and 1933, Kermack and McKendrick published a series of three papers [29, 30, 31] that laid the foundations of two of the most widely used models in epidemiology, the SIR and the SIS model (for *susceptible - infected - removed* and *susceptible - infected - susceptible* respectively). They first assumed a closed system, in which initially the individuals are susceptible of becoming ill, except for a fraction which is already infected. If an individual gets infected, he can in turn spread the illness, until he is removed either by death or by recovery conferred by a permanent immunity (SIR model, [29]). Then, they introduced population dynamics at the birth level (but the only cause of death is the illness still), and supposed that the recovered individuals may become ill again (the SIS model is a special case, see [30]). Finally, they included a non-specific death rate, independent of the illness, to account for other causes of death [31].

The complexity of the models analysed is twofold: on one hand, the rates (infectivity, recovery, birth, etc.) can be either constant or functions, and on the other hand, more transitions are allowed (non-fatal outcome of an illness, birth and death rates of the individuals). In spite of this increasing complexity in the mathematical analysis of the models, it is remarkable that they drew very similar conclusions. First of all, a threshold density of population was found, in order for an epidemic to take place: below that level, no epidemic can occur. Second, in most cases steady states were deduced, when population dynamics are taken into consideration. Finally, these steady states were found to be stable in most cases, but when they are not there are two possible outcomes: either the epidemic wanes, either the population is wiped out.

Information retrieval

One of the key goals that has motivated a large amount of work in the field of information retrieval[3] is that of *dimensionality reduction*: how to correctly describe a set of text documents, without having to store, for each of them, all of their words one by one as a in raw form? This is an important issue, since dimensionality reduction allows a simplified treatment of large amounts of text. As seen in fore coming chapters, a dimensionality reduction technique, the latent Dirichlet allocation (LDA, [17]), is applied to text documents in the context of this work, which are modeled as LDA vectors. Other state variables of the users, are also represented as LDA vectors. Therefore, a brief history of information retrieval, which lead to the development of LDA, is mentioned.

In this respect, one of the earliest solutions to the previous problem was the *tf-idf* scheme. First, a fixed vocabulary is chosen, and then for each document the product of the normalized frequency by the logarithm of the normalized specificity is computed for each word of the vocabulary. Therefore, the tf-idf score of a document for a given word is not only a function of its frequency in the document, but also of how well it enables to identify that document in particular instead of another, which is suitable behavior. This way, the dimensionality reduction achieved consists in the description of each document by a vector of fixed length (equal to the size of the vocabulary), instead of a list containing all of the words in it, and a term-by-document matrix X is obtained.

However, soon the dimensionality reduction achieved with tf-idf was deemed to be insufficient, as the size of the vocabulary could still be considerable. In order to further reduce the X matrix, the *latent semantic indexing* (LSI) was proposed, which "uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that captures most of the variance in the collection" [17, p. 994]. With LSI, the problem of dimensionality reduction was solved quite well, as X was now a topic-by-document matrix where the number of topics is comparatively small, each of them being a linear combination of the terms chosen in tf-idf. In other words, each document is now represented by a vector of topic weights, instead of vector of terms weights.

Nevertheless, the LSI model does not propose a probabilistic mechanism for document generation. In order to fulfill this gap, the *probabilistic latent semantic indexing* (pLSI) was created. Now, each word of a document is sampled from a topic, defined as a probability distribution over a fixed vocabulary: thus, pLSI allows that words from the same document may belong to different topics. Although the pLSI is a probabilistic model, no generation mechanism is provided for the topics, which yields two drawbacks: first, it is not clear how to assign the topic probabilities for a document not belonging to the training set, and second, the number of parameters of the probabilistic model grows linearly with the number of documents.

[3]This account of past research in information retrieval is based on Blei et al. [17, p. 993–995].

Finally, the *latent Dirichlet allocation* (LDA) — a probabilistic model as well — was proposed. In the LDA scheme, "each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities" [17, p. 993]. The topics are then generated from a probabilistic distribution (a Dirichlet distribution indeed) whose parameters are the same across the documents. Therefore, the number of parameters ceases to grow linearly with the number of documents, despite it still grows linearly with the size of the vocabulary. With LDA, a document is represented as a vector of probabilities over the underlying set of topics, as with pLSI.

Perceptual choice

The model of information diffusion presented in this work involves decisions of the OSN users, such as the election of a thread for browsing, or of a post for reading. In order to simulate those decisions, a perceptual choice model, the leaky competing accumulator (LCA, [44]) is used in the main model. Below, a review of perceptual choice research is performed[4], and the LCA model is introduced.

The field of perceptual choice studies deals with the problem of understanding how the brain decides between two or more alternatives, based on sensory perceptions. For example, if a subject is asked for whether a light spot is green or red, what are the underlying mechanisms that are at work before giving an answer? In particular, two attributes of the answer are of interest: the choice that has been made, and the time that the subject needed in order to decide. Evidence has shown that both the decision and the duration tend to be quite variable [44, p. 550]. Thus, models of perceptual choice must account for these and other empirical findings.

Previous work in the field revolves around two principles. On the one hand, "they treat information processing as a gradual process, based on the accumulation of information over time" [44, p. 550]. On the other hand, "they treat the process as stochastic or intrinsically variable, so that the information accumulated within each small time interval is subject to random fluctuations" [44, p. 550]. Two basic kinds of models will be discussed here: accumulator models (or counter models) and random walk models.

In the case of the accumulator models, the underlying process is analogous to sampling balls (red and green) from an urn, with replacement. Two counters (one for each color) are initialized with a zero value, and each time that a ball is drawn the corresponding counter is incremented. When a counter reaches a criterion value, the related decision is taken [44, p. 551]. As Usher and McClelland note, "increments may be binary, multivalued, or continuous, and sampling may be assumed to occur at discrete time steps or continuously" [44, p. 551].

[4]The following account of past research in perceptual choice is based on Usher and McClelland [44, p. 550–557].

The accumulator models reflects well the two principles mentioned above, as the decision is taken after a gradual process during which information is accumulated, and the accumulation itself is stochastic.

The next class of perceptual choice models are random walk models. The main difference with accumulator models is that a single variable is now considered, which is the difference of evidence between two alternatives. Recalling the previous analogy, this time a single counter is initialized with a zero value, and each time that a ball is drawn, the counter is either increased or decreased, according to the color, until a criterion value is reached. [44, p. 551]. In a similar fashion, increments may be either discrete or continuous, and time may be sampled either in a discrete or a continuous way (in the former case, it is a *diffusion process*). An important type of diffusion process is *classical diffusion processes*, in which the increment is sampled from two continuous distributions with the same standard deviation. For example, if a green ball is drawn, the increment is sampled from a normal distribution with mean $\mu_g > 0$ and standard deviation σ; otherwise, if the ball drawn is red, the increment is sampled from a normal distribution with mean $\mu_r < 0$ and equal standard deviation σ.

A problem from which suffer both accumulator and random walk models, is that of perfect accuracy: the alternative with the greatest level of sensory evidence always wins. But it is known that humans, for example, do not exhibit such perfect accuracy, but rather reach an *accuracy ceiling*, even if unlimited time is allowed before giving an answer [44, p. 551]. In order to account for this issue, a variant of the classical diffusion process has been proposed, where the means of the increments are themselves random variables with variance: that is, *diffusion-with-drift-variance* models [44, p. 552]. Other suggested possibility is that the subjects can sample only a finite number of useful observations [44, p. 552]. Another problem faced, in particular by random walk models, is how to extend the analysis for the general case of N alternatives.

Usher and McClelland therefore incorporated two additional information principles, based on the previous work's drawbacks and on empirical evidence. First, they assumed that information accumulation is subject to leakage or decay [44, p. 552–553,555]. Second, representations of the alternative outcomes of the decision process compete with each other, through a process of lateral inhibition [44, p. 555-556]. Indeed, these two additional principles "nicely dovetail with neurophysiological evidence that is considered below, suggesting that such mechanisms are indeed at work in the neural machinery underlying performance in information processing tasks" [44, p. 553]. By considering two more principles, self-excitation and nonlinearity [44, p. 553], Usher and McClelland finally formulated the *leaky competing accumulator* (LCA) model [44]) which consists in a set of N stochastic differential equations (one for each alternative), that model the information principles (information accumulation, stochastic noise, leakage, competition through lateral inhibition, self-excitation, nonlinearity) mentioned above.

1.2 Research hypothesis

The fundamental intuition underlying this work is that it possible to combine techniques and models arising from General Diffusion, Decision Making and Text Mining, in order to study properly the diffusion of information through a social network. Specifically, two research hypotheses are formulated:

R.H. 1 (*superiority of LCA as the decision mechanism*) the LCA model, as the underlying mechanism of decision, gives the best prediction of actual contents and graph generation, in comparison with a voter model, a Logit-based model and a deterministic model.

R.H. 2 (*decay of contents variance over time*) the variance of the LDA profiles across users decreases with time.

1.3 Thesis objectives

1.3.1 General objective

The main goal of this research work is to create, implement and evaluate an information diffusion model through a social network.

1.3.2 Specific objectives

S.O. 1 To perform a review of literature in general diffusion, decision making and text mining.

S.O. 2 To create a model of LDA topics diffusion through a social network.

S.O. 3 To implement the model of LDA topics diffusion through a social network.

S.O. 4 To evaluate and validate the quality of the obtained results, comparing with existing algorithms.

S.O. 5 To evaluate the practical usefulness of the model developed.

1.4 Expected results

E.R. 1 (S.O. 1) Chapter 2, 3 and 4 of this report.

E.R. 2 (S.O. 2&3) Library written in JAVA, containing the code developed and the auxiliary libraries needed.

E.R. 3 (S.O. 4) Chapter 7 of this report.

E.R. 4 (S.O. 5) Chapter 7 of this report.

General objective			
To develop, implement and evaluate an information diffusion model through a social network			
Specific objective 1	**Specific objective 2 & 3**	**Specific objective 4**	**Specific objective 5**
To perform a review of literature in general diffusion, decision making and text mining	To create and implement a model of LDA topics diffusion through a social network.	To evaluate and validate the quality of the obtained results, comparing with existing algorithms	To evaluate the practical usefulness of the model developed
Expected result 1	**Expected result 2**	**Expected result 3**	**Expected result 4**
Chapters 2, 3 and 4 of this report	Library written in JAVA, containing the code developed and the auxiliary libraries needed	Chapter 7 of this report	Chapter 7 of this report

Table 1.2: General objectives, specific objectives and expected results of this work.

1.5 Structure of this report

In chapters 2, 3 and 4, a theoretical framework is built, in order to give sense to the model and the results obtained. In chapter 5, the main model implemented, as well as the benchmark models, are discussed. In chapter 6, the methodological framework is explained, focusing on how the experiments are carried out, and how their results are evaluated. Then in chapter 7, the results obtained by simulation are presented. Finally, in chapter 8, the conclusions are drawn, and some possibilities for future work explored.

Chapter 2

Diffusion: General Background

The diffusion problem is found in the literature under many forms. It may be a pathogen spreading through a human or animal population, a computer virus among a network of computers, a new product among consumers, a candidate decision among voters, a new word among the speakers of a language, only to name a few examples. Diffusion research is therefore a point of interchange between various scientific traditions, such as sociology, physics, epidemiology, graph theory, stochastic processes, chemistry, computer science, etc. One of the appealing features is that in spite of various motivations and origins of the problems, it seems that the mathematical formalism to which are converging the previous traditions is in general the same. The network is a standard representation of the environment, even in continuous cases. For example, a heat transfer process is a diffusion process on a continuous medium, but is in general solved within a mesh of finite elements, whose contacts define a network. The diffusion mechanism may be either deterministic (as in the case of heat transfer), or stochastic (as in the case of the spread of an epidemic). In fact, the *modeled* diffusion mechanism may be stochastic, reflecting the impossibility of complete knowledge and measurement of the underlying phenomena.

The importance of the network structure has been recognized in the second half of the 20^{th} century. Nonetheless, due to time restrictions and to this work's author skills, it has been chosen to review the very basics of diffusion research[1]. Seminal publications in sociology, epidemiology and marketing have been selected and explained, in order to build the best possible foundations for future work. These publications all assume a continuous transmission environment, and a network equivalent is deduced at the end of this chapter. In addition, a model used in interacting particle systems, the voter model, is presented, which do operate on network (a regular lattice indeed). This model can be generalized, and it has been in fact. At the end of each section, some links are established with this work's practical experiments, in an attempt to give some theoretical support to the obtained results.

[1]More complete reviews are found in [24, Chapter 19 and 21] and [28, Chapter 7].

2.1 The Diffusion of Innovations

The diffusion of innovations is a line of research in sociology rooted in the beginning of the 20^{th} century, although it became mainstream during the second half. It is interesting to note that rural sociology made significant contributions to the field, in particular in its beginnings. The main concern of diffusion of innovations research is to understand why certain innovations successfully propagate, while others don't [41, p. 140][37, p. 1]. Below, three important works are discussed. The first is *The Laws of Imitation* by Gabriel Tarde, which may be labeled as the starting point in diffusion of innovations research. Then comes the *Diffusion of Innovations* of Everett Rogers, which brought a wider attention among the scientific community. Finally, a classic mathematical model, the Bass' model, is presented.

2.1.1 The importance of imitation

The role of imitation in the diffusion of ideas in human societies is discussed at length by French sociologist Gabriel Tarde, in his seminal book *The Laws of Imitation* [41]. Years after the first publication of his book in 1890, it remained highly influential and was cited by scholars in the fields of the diffusion of innovations and social networks analysis. For example, Everett Rogers affirms:

> "The intellectual tradition that we refer to as "early sociology" traces its ancestry to a French sociologist, Gabriel Tarde (1903), but most of the research publications in this tradition appeared from the late 1920s to the early 1940s. The true significance of the field lies not in its volume of investigations nor in the sophistication of its research methods but in the considerable influence of early sociologists upon later diffusion researchers.
>
> Tarde (1903) proposed several novel notions for testing by later diffusion investigators. He was among the first to suggest that the adoption of a new idea follows a normal, S-shaped distribution over time. He also argued that the greater cosmopoliteness of innovators is one reason for their early adoption of new ideas. Probably Tarde's greatest contribution was his insight into the process by which the behavior of opinion leaders is imitated by other individuals." [37, p. 52]

Moreover, the opinion of the author of the present work is that the *Communication of innovations* of Rogers [37] is based largely on Tarde's ideas, presenting a unified framework as shown in the next subsection. On the other hand, economist Matthew O. Jackson underlines the significance of Tarde's work, concerning the conceptualization of the S-shaped curve and the role of imitation in the diffusion of innovations [28, p. 242]. Therefore, it is pertinent to review some of Tarde's insights below. Four topics will be of interest: the role of imitation, opinion leaders, cosmopoliteness and the S-shaped curve.

The role of imitation in diffusion

According to Tarde's account, imitation plays a central role in the formation and consolidation of social ties. There cannot be real social ties without imitation between the concerned individuals, and the very act of refusing to imitate implies an *antisocial* relationship [41, p. XIX]. Effective imitation requires a common background of shared knowledge, customs, education, etc, as imitation consists in the repetition of long learned habits. It was postulated that a society is defined by a set of common goals, and that the social link emerges with the need of the other. But Tarde denies this view, and affirms that a relationship based on similarity is much deeper:

> "Social relations, I repeat, are much closer between individuals who resemble each other in occupation and education, even if they are competitors, than between those who stand the most in need of each other. Lawyers, journalists, magistrates, all professional men, are cases in point. So society has been properly defined by common speech as a group of people who, although they may disagree in ideas and sentiments, yet, having had the same kind of bringing up, have a common meeting ground and see and influence one another for pleasure." [41, p. 64]

Thus, "society is imitation" [41, p. 87], and every social fact has imitation as its cause [41, pp. 14–15 and 50]. Imitation is the mechanism that enables the *universal repetition* [41, p. 1], which can be thought of as a propagating wave which can be traced back to its origin. The analogy with undulation is explicitly stated in the following quotation for example:

> "Men of different civilisations come into mutual contact on their respective frontiers, where, independently of war or trade, they are naturally inclined to imitate one another. And so, without its being necessary to displace one another in the sense of checking the spread of one another's examples, they continually and over unlimited distances react upon one another, just as the molecules of the sea drive forward its waves without displacing one another in their direction." [41, p. 49]

Generalizing this view, human activity may be seen as "*individual initiative followed by imitation*" [41, p. 3], where invention and imitation are the elementary social acts [41, p. 144], the former propagating through the latter:

> "In brief, the picture of primitive society which rises before me is that of a feeble, wayward imagination scattered here and there in the midst of a vast passive *imitativeness* which receives and perpetuates all its vagaries as the water of a lake circles out under the stroke of a bird's wing on its surface." [41, p. 95]

Opinion leaders

Tarde studies the principle of the *imitation of the superior by the inferior* [41, p. 213], which postulates that imitation proceeds as a *descent* [41, p. 214], as can be read in the following:

> "The principal role of a nobility, its distinguishing mark, is its initiative, if not inventive, character. Invention can start from the lowest ranks of the people, but its extension depends upon the existence of some lofty social elevation, a kind of social *water-tower*, whence a continuous water-fall of imitation may descend. At every period and in every country the aristocratic body has been open to foreign novelties and has been quick to import them [...]" [41, p. 221]

Therefore the highest social classes of distant places tend to resemble each other, although the local populations in general still keep their differences [41, p. 220]. The elite of a society then import innovations from the outside, and may generalize innovations from the lower classes, because of the natural human tendency into imitating the hierarchical superior [41, p. 218]. Imitation also acts in the opposite sense, but in a lesser extent [41, p. 215].

An important remark is that the difference between the imitator and the imitated must not be too great, or no imitation can really occur [41, p. 224]. In this respect, Tarde introduces the notion of *social distance*:

> "[...] in reality, the thing that is most imitated is the most superior one *of those that are nearest*. In fact, the influence of the model's example is efficacious inversely to its *distance* as well as directly to its superiority. *Distance* is understood here in its sociological meaning. However distant in space a stranger may be, he is close by, from this point of view, if we have numerous and daily relations with him and if we have ever facility to satisfy our desire to imitate him. "[41, p. 224]

Similarly, Tarde notes that there exists an optimal point of the degree of communication that maximizes imitation [41, p. 392]. A very similar argument is advanced by Rogers, concerning heterophily [37, p. 15]. As a final remark, the notion of social distance may be thought of as anticipating the posterior emergence of Social Network Analysis, with the graph as the main tool in the study of social networks.

Cosmopoliteness and the interplay between the traditional and the modern

A fundamental distinction has to be made between *custom* and *fashion* [41, p. 244]. Custom is made of immutable beliefs and values, that have long passed with success the test of time, and therefore constitute the very foundations of the morality of a society [41, p. 247]. Custom is taught to an individual since early childhood, by the means of paternal authority, and consists of ancestral knowledge [41, pp. 245–246]. Therefore, custom is related to obedience, past times and geographical proximity. On the contrary, fashion is by nature much more volatile. A society's member is usually exposed to fashion at a grown-up age, under a foreign influence of some sort. Fashion is related to free choice[2], focus on present times and on geographical remoteness. It appears therefore that fashion is but a superficial perturbation of custom:

[2]Free choice only in appearance, as the openness to new ideas obeys to an old mental structure which was built by custom [41, p. 246]

> "Imitation, then, that is engaged in the currents of fashion is but a very feeble stream compared with the great torrent of custom. And this must be necessarily so. But, however slender this stream may be, its work of inundation or irrigation is considerable, and it behoves to us to study its periodic rises and falls in the very irregular kind of rhythm in which they occur." [41, p. 244–245]

In this context, cosmopoliteness can be seen as the degree of exposition of an individual to external influences, in part through fashion. From the paragraph above, it must be noted that the cosmpolite individual does not lack necessarily of custom, and indeed he does not. But he has the ability to comprehend new forms of communication (thus of imitation) that do not exist in the traditional frame of mind to which he belongs. Therefore, it is easy to understand why cosmopoliteness can play an important part in the diffusion of innovations:

> "Primitive rural communities can only imitate their fathers, and so they acquire the habit of ever turning towards the past, because the only period of their life in which they are open to the impressions of a model is their infancy, the age that is characterised by nervous susceptibility, and because, as children, they are under paternal rule. On the other hand, the nervous plasticity and openness to impressions of adults in cities is in general well enough preserved to permit them to continue to model themselves upon new types brought in from outside." [41, pp. 247–248]

The S-shaped curve

In the first place, Tarde wonders why, of all the possible civilisations that could have existed in Europe at the time, only one observable kind of civilisation (rooted in the antique greco-roman heritage) exists, and not another. Acknowledging that reality is only one realisation of infinitely many possibilities, Tarde concludes that all past civilisations indeed *claimed to universality* [41, p. 22], but those who failed did so because of conflict. Generalizing the argument, he argues that an innovation tends to follow a geometric progression [41, pp. 22 and 115] indefinitely. But after some time, an innovation may clash with another, and its progression may decline, or even stop at all [41, pp. 115–116]. In other words, an innovation starts slowly, and may begin to rise because of its natural tendency to propagate, with imitation being the underlying mechanism at work, until it begins to wane under the competitive pressure exerted by another innovation. The result is an S-shaped curve, best described in the following quotation:

> "A slow advance in the beginning, followed by a rapid and uniformly accelerated progress, followed again by progress that continues to slacken until it finally stops: these, then, are the three ages of those real social beings which I call inventions or discoveries. None of them is exempt from this experience any more than a living being from an analogous, or, rather, identical, necessity. A slight incline, a relatively sharp rise, and then a fresh modification of the slope until the plateau is reached: this is also, in abridgment, the profile of every hill, its characteristic curve." [41, p. 127]

The process is not necessarily that clean, and the progression of an innovation can exhibit various cycles of rise and decline, given that interactions with other innovations are not always destructive, but may also be creative:

> "[...] we may be sure, upon inspecting a given curve, particularly if it has been plotted according to the rules that were given some pages back, that as soon as the first obstacles are overcome and it has assumed a well-marked upward movement according to a definite angle, every upward deviation will reveal the insertion of some auxiliary discovery or improvement at the corresponding date, and every drop towards the horizontal will reveal, on the other hand, according to our foregoing law, the shock of some hostile invention." [41, p. 129]

Another possibility that accounts for the slowdown of the propagation of an innovation may also simply be the depletion of the propagating environment: there is no more space left to grow.

Final remarks

Three great contributions of Tarde must be stressed. On the one hand, he emphasises the *social* as a legitimate category of natural phenomena, as occurs with the chemical, the physical, the biological, and so on. The social may be rooted in psychological and neurological mechanisms, but it is perfectly possible to study social phenomena in their own, abstracting lower layers. Tarde, in this sense, anticipates the great deal of social measurement that characterizes the 20^{th} century. On the other hand, he places imitation as the main "force" of social events, putting on a second plane the role of utility. In this context, users of a social network would copy one another with the sole objective of imitation, with no utility maximization and the like intended. This represents a shift from the traditional economic model of agents, traditionally used in marketing. Imitation is the starting point and not a consequence that must be explained. Finally, the importance that Tarde assigns to the elites, and the superior in general, which he connects to cosmopoliteness, may be interpreted at a smaller scale in a OSN. The elites in this case would be well connected users, *i.e.* key members. Around key members emerge communities, characterized by stronger social ties than with the outside, which is equivalent to a deeper level of imitation and similarity.

2.1.2 A unified framework

Before discussing the framework built by Rogers[3], it must be noted that more than seventy years passed between the first publication of Tarde's *Laws of Imitation* [41, 42] and the first publication of Rogers' *Diffusion of Innovations* [36, 37]. Although no review of diffusion research has been made by the author of this work in that intermediary period, two influ-

[3]This discussion is based on the chapters 1 and 2 of Rogers, *Communication of Innovations: a Cross-Cultural Approach* [37]

ential studies will be mentioned. The first one is the study of the diffusion of hybrid seed corn in Iowa by Ryan and Gross [39], which concluded that "commercial channels, especially salesmen, were most important as original sources of knowledge, while neighbors were most important as influences leading to acceptance" [39, p. 15]. The second one is the study of the diffusion of a new drug among physicians by Coleman et al. [20], where it was found that professional ties operate first, followed by friendship ties, thus revealing the importance of the network structure [20, p. 268].

As Rogers points out, diffusion research is a special type of the broader field of communication research (figure 2.1). As the message is new in the case of diffusion, its adoption conveys a degree of risk, resulting in a different behavior compared to the adoption of routine ideas [37, pp. 12–13]. Moreover, the focus is on *overt behavior change*, rather than changes in knowledge or attitudes [37, p. 13], and as Rogers states: "the knowledge and persuasion effects of diffusion campaigns are considered mainly as intermediate steps in an individual's decision-making process leading eventually to overt behavior change" [37, p. 13]. Two aspects of diffusion research will be considered: social change, and the diffusion of innovations itself.

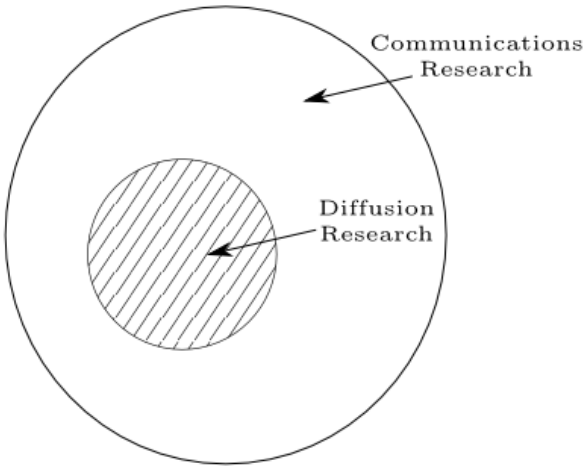


Figure 2.1: "Diffusion research is that subset of communication research dealing with the transfer of new ideas" [37, p. 12]. *Source: reproduced from [37, p. 12]*

Social change

Social change is defined as "the process by which alteration occurs in the structure and function of a social system" [37, p. 7]. The source of social change may be either internal or external to a social system; the former case corresponds to *immanent change*, while the latter is labeled *contact change* [37, p. 8]. Further, a distinction may be made according to the recognition of the need for change; when the recognition happens inside the social system, then *immanent change* or *selective contact change* occurs, otherwise it is *induced immanent*

change or *directed contact change* [37, pp. 8–9]. In the diffusion research tradition, most studies are focused on directed contact change [37, p. 10]. On the other hand, social change occurs at two levels of analysis: the *individual* level (microanalytic approach) and the *social system* level (macroanalytic approach) [37, pp. 10–11]. The previous levels are not separated but, rather, interdependent: "change at these two levels is closely interrelated. [...] the aggregation of a multitude of individual changes produces a system-level alteration" [37, p. 11].

The diffusion of innovations

The diffusion of an innovation consists of "(1) the *innovation* (2) which is *communicated* through certain *channels* (3) over *time* (4) among the members of a *social system*" [37, p. 18]. As it deals with the diffusion of new ideas or products, the time dimension is particularly relevant [37, p. 18]. The innovation, the communication channels, the time and the social system are readily discussed below.

Rogers defines the innovation as follows:

> "An *innovation* is an idea, practice, or object perceived as new by an individual. It matters little, so far as human behavior is concerned, whether or not an idea is "objectively" new as measured by the lapse of time since its first use or discovery. It is the perceived or subjective newness of the idea for the individual that determines his reaction to it. If the idea seems new to the individual, it is an innovation." [37, p. 19]

Moreover, in accordance with Tarde's thought, "every idea has been an innovation sometime" [37, p. 19]. An innovation has always an *idea* component, but not necessarily an *object* component [37, p. 21]. The adoption of the idea is called a *symbolic* decision, while the acceptance of the object is named an *action* decision [37, p. 21]. Rogers lists five characteristics of innovations that contribute to their adoption:

1. *relative advantage:* perceived superiority of the innovation [37, p. 22]
2. *compatibility:* "with the existing values, past experiences, and needs of the receivers" [37, p. 22]
3. *complexity:* difficulty of comprehension and use [37, p. 22–23]
4. *trialability:* "degree to which an innovation may be experimented with on a limited basis" [37, p. 23]
5. *observability:* "degree to which the results of an innovation are visible to others" [37, p. 23]

The previous list is not complete and only shows the most important characteristics of an innovation [37, p. 23].

Communication is defined as "the process by which messages are transmitted from a source to a receiver. In other words communication is the transfer of ideas from a source with a viewpoint of modifying the behavior of receivers. A communication *channel* is the means by which the messages gets from the source to the receiver" [37, p. 24]. The process, in turn, can be divided as follows:

> "We might think of the communication process in terms of the oversimplified but useful S-M-C-R model. A *source* (S) sends a *message* (M) via certain *channels* (C) to the *receiving* individual (R)." [37, p. 11]

It can be noted that the diffusion process — consisting of the sequence innovation, communication, time, social system — is similar to the communication process. This makes sense since diffusion research originates from communication research [37, pp. 18–19 and 24]. Indeed, the innovation component in the decision process is the equivalent of the message component in the communication process; the same holds for the social system and the receivers; finally, both the diffusion process and the communication process include a channel component. The difference lays in the emphasis that the diffusion process puts on time.

Now, time is an important element of the diffusion process [37, p. 24], and Rogers states: "The time dimension is involved (1) in the innovation-decision process by which an individual passes from first knowledge of the innovation through its adoption or rejection, (2) in the innovativeness of the individual, that is, the relative earliness-lateness with which an individual adopts an innovation when compared with other members of his social system, and (3) in the innovation's rate of adoption in a social system, usually measured as the number of members of the system that adopt the innovation in a given time period" [37, pp. 24–25]. The innovation-decision process is conceptualized into four steps, each one of them being related to a corresponding function [37, p. 25]:

1. *knowledge* (knowledge function)
2. *persuasion* (persuasion function)
3. *decision* (decision function)
4. *confirmation* (confirmation function)

The innovativeness is defined as "the degree to which an individual is relatively earlier (in terms of actual time of adoption) in adopting new ideas than the other members of his system" [37, p. 27]. Members of a social system are furthermore divided into five *adopter categories*, in decreasing order of innovativeness: (1) innovators, (2) early adopters, (3) early majority, (4) late majority, (5) laggards [37, p. 27]. The time involved in the innovation-decision process explains the time lags observed in the adoption of innovations [37, pp. 16–17]. Finally, the *rate of adoption* is "the relative speed with which an innovation is adopted by members of a social system" [37, pp. 27–28], and is "usually measured by the length of time required for a certain percentage of the members of a social system to adopt an innovation" [37, p. 28].

Concerning the last element of the diffusion of a new idea, "a *social system* is defined as a collectivity of units which are functionally differentiated and engaged in joint problem solving with respect to a common goal. The members or units of a social system may be individuals, informal groups, complex organizations, or subsystems"[4] [37, p. 28]. In turn, the social system exhibits a *social structure*, which exists "to the extent that the members in a social system are differentiated" [37, p. 28] and develops "through the arrangement (such as in an hierarchical fashion) of the statuses or positions in a system" [37, p. 29]. Both diffusion and social structure are interrelated [37, p. 29], and the structure fosters or inhibits the diffusion [37, p. 29] as well as the diffusion changes the structure [37, p. 30]. In addition to its social structure, a social system also has *norms*, which are "the established behavior patterns for the members of a given social system" [37, pp. 30–31]; they also "define a range of tolerable behavior and serve as a guide or a standard for the members of a social system" [37, p. 31]. Actual norms are located in a continuum between two ideal types: *traditional norms* and *modern norms* [37, p. 33]. Given the importance of norms for individual innovativeness [37, pp. 29–30], these two ideal types will be described. On the one hand, "traditional social systems can be characterized by:

1. Lack of favorable orientation to change.

2. A less developed or "simpler" technology.

3. A relatively low level of literacy, education, and understanding of the scientific method.

4. A social enforcement of the status quo in the social system, facilitated by affective personal relationships, such as friendliness and hospitality, which are highly valued as ends in themselves.

5. Little communication by members of the social system with outsiders. Lack of transportation facilities and communication with the larger society reinforces the tendency of individuals in a traditional system to remain relatively isolated.

6. Lack of ability to empathize or to see oneself in others' roles, particularly the roles of outsiders to the system. An individual member in a system with traditional norms is not likely to recognize or learn new social relationships involving himself; he usually plays only one role and never learns others." [37, p. 32]

On the other hand, "a modern social system is typified by:

1. A generally positive attitude toward change.

2. A well developed technology with a complex division of labor.

3. A high value on education and science.

4. Rational and businesslike social relationships rather than emotional and affective.

[4]Tarde would say that a social system is defined as a collectivity of units which imitate one another (section 2.1.1)

5. Cosmpolite perspectives, in that members of the system often interact with outsiders, facilitating the entrance of new ideas into the social system.

6. Empathic ability on the part of the system's members, who are able to see themselves in roles quite different from their own." [37, p. 32–33]

Finally, one role of interest in a social system is the *opinion leader* [37, p. 34], where "*opinion leadership* is the degree to which an individual is able to informally influence other individuals' attitudes or overt behavior in a desired way with relative frequency" [37, p. 35].

Traditions of diffusion research

Rogers distinguishes eight traditions of research in innovations diffusion research:

anthropology: "the anthropology diffusion tradition is the oldest of the seven traditions. It has had great influence on the early sociology, rural sociology, and medical sociology fields but only limited impact on the other traditions" [37, p. 48]

early sociology: already mentioned above, see section 2.1.1

rural sociology: "the research tradition which boasts the largest and most enduring concern with diffusion is rural sociology" [37, p. 53]

education: "one of the larger traditions in terms of the number of studies, education is one of the lesser traditions in terms of its contribution to understanding the diffusion of innovations or to a theory of social change" [37, p. 58]

medical sociology: "the innovations studied have consisted of (1) either new drugs or medical techniques, where the adopters are doctors, or (2) polio vaccine, family planning methods, or other medical innovations, where the adopters are clients or patients" [37, p. 62]

communication: "one of the important concerns of these communication researchers is the diffusion of news events carried by the mass media" [37, p. 67]

marketing: "marketing managers of firms in the U.S. have long been concerned with how to launch new products most efficiently. Their interest in this topic is sparked by the appearance of large numbers of new consumer products and by the high rate of failure of such products" [37, p. 68]

other traditions: include *agricultural economics*, *geography*, *general economics*, *speech*, *general sociology*, and *psychology* [37, p. 69–71]

Final remarks

The Communication of Diffusions [37] of Rogers performs a complete review of sociological research on the diffusion of innovations until the 1960s (in the case of the second edition). Rogers presents relevant aspects in a succinct way, with many references and examples, and builds up a framework still useful today. The most important contribution of Rogers, in the scope of this work, is the connection with sociological concepts it enables. Though technical in nature, the models implemented in this work still try to simulate the behavior of human beings and communities. Therefore, it is mandatory to provide, however little this may be, a sociological basis for the behavior of humans in the context of the diffusion of information. Though the diffusion of innovations is more restrictive in this sense, as innovations are new information, it nonetheless provides a useful framework of analysis. In particular, the concept of an optimum heterophily for diffusion is very interesting.A different approach is undertaken in this work, as the simulated agents tend to maximize the similarity with their own preferences in their choices.

2.1.3 The Bass Model

In 1969, Frank M. Bass published his paper *A new product growth model for consumer durables* [16], which presents a model for the sales of a new product as a function of time. Mathematically, the theory originates from contagion models, and, in more conceptual terms, is inspired on previous research about the diffusion of innovations (in particular on Rogers' work). The goal of Bass' model is to predict the timing of adoption of a new product (its purchase time).

As mentioned above, Rogers distinguishes five adopter categories: (1) innovators, (2) early adopters, (3) early majority, (4) late majority and (5) laggards (section 2.1.2). Bass aggregates categories (2) to (5) into a single one: *imitators* [16, p. 216]. As imitators are influenced by their fellows, the more a new product is being purchased, the more an imitator will experiment a purchase pressure; on the contrary, innovators are insensitive to the purchases made by other members of a social system [16, p. 216]. This can be stated as follows: "The probability that an initial purchase will be made at T given that no purchase has yet been made is a linear function of the number of previous buyers" [16, p. 216]. In mathematical terms:

$$P(T) = p + \frac{q}{m} Y(T) \tag{2.1}$$

$P(T)$ is the probability that an initial purchase will be made at T given that no purchase has been made, p is the fraction of all adopters who are innovators (the coefficient of innovation), q is the fraction of all adopters who are imitators (the coefficient of imitation), m is the total number of initial purchases spanning the life time of the product (market size) and $Y(T)$ is the number of previous buyers.

Continuous time case

Supposing the the purchases occur in continuous time, let $f(T)$ and $F(T)$ be the density and distribution function respectively of the time of initial purchase. By definition, $F(T)$ is the probability that an initial purchase happened before T, and therefore $1 - F(T)$ is the probability that an initial purchase has not occurred yet at T. Therefore, equation 2.1 may be rewritten as:

$$\frac{f(T)}{1 - F(T)} = p + \frac{q}{m} Y(T) \tag{2.2}$$

Let $S(T)$ be the sales performed at T, then $S(T) = mf(T)$. Moreover, $Y(T) = mF(T)$, and equation 2.2 becomes:

$$\frac{f(T)}{1 - F(T)} = p + qF(T) \tag{2.3}$$

And as $f = \mathrm{d}F/\mathrm{d}T$, then Bass' differential equation is:

$$\frac{\mathrm{d}F(T)}{\mathrm{d}T} = [1 - F(T)][p + qF(T)] \tag{2.4}$$

Given the initial condition $F(0) = 0$, the solution of equation 2.4 is:

$$F(T) = \frac{1 - \mathrm{e}^{-(p+q)T}}{\frac{q}{p}\mathrm{e}^{-(p+q)T} + 1} \tag{2.5}$$

And the density function is:

$$f(T) = \frac{(p+q)^2}{p} \frac{\mathrm{e}^{-(p+q)T}}{(\frac{q}{p}\mathrm{e}^{-(p+q)T} + 1)^2} \tag{2.6}$$

Deriving $f(\cdot)$:

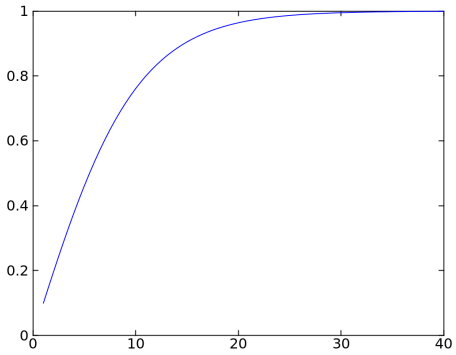
$$\frac{\mathrm{d}f(T)}{\mathrm{d}T} = \frac{\mathrm{d}^2F(T)}{\mathrm{d}T^2} = \frac{(p+q)^3}{p}\mathrm{e}^{-(p+q)T} \frac{\frac{q}{p}\mathrm{e}^{-(p+q)T} - 1}{(\frac{q}{p}\mathrm{e}^{-(p+q)T} + 1)^3} \tag{2.7}$$

If $p \geq q$, then $\mathrm{d}f(T)/\mathrm{d}T \leq 0 \quad \forall T \geq 0$, therefore $f(\cdot)$ is decreasing and $F(\cdot)$ is concave. If $p < q$, then $\exists T^* = \frac{\ln(q/p)}{p+q}$ such that $f(T^*) = 0$. Thus, T^* is the time at which sales reach their peak and at which $F(\cdot)$ changes of concavity: $F(\cdot)$ is S-shaped (figure 2.2).

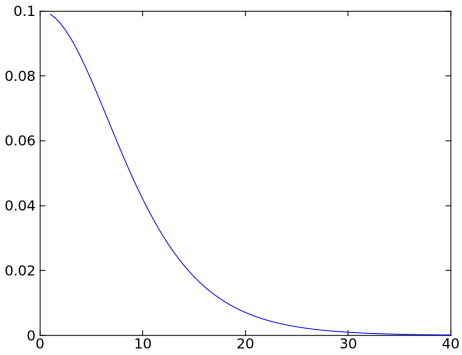
Discrete time case

Supposing now that time passes with discrete steps, let f_T be the probability that a purchase is made at T, and F_T the probability that a purchase has been done before or at T. The discrete equivalent of equation 2.3 is then:

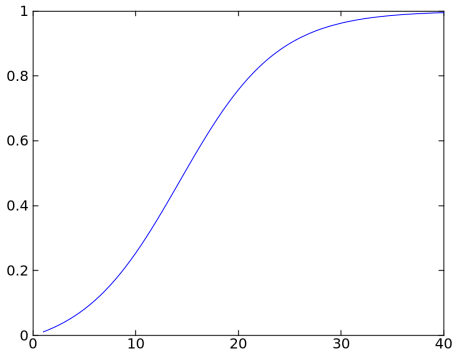
$$\frac{f_T}{1 - F_{T-1}} = p + qF_{T-1} \tag{2.8}$$



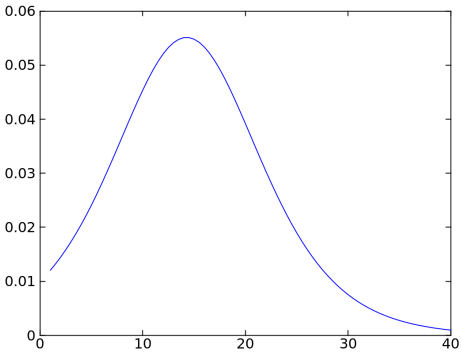
(a) $F(T)$ ($p = 0.1$ and $q = 0.1$)



(b) $f(T)$ ($p = 0.1$ and $q = 0.1$)



(c) $F(T)$ ($p = 0.01$ and $q = 0.2$)



(d) $f(T)$ ($p = 0.01$ and $q = 0.2$)

Figure 2.2: (a) and (b): $p \geq q$ case, with $p = q = 0.1$. **(c) and (d)**: $p < q$ case, with $p = 0.01$ and $q = 0.2$. Note that when $p \geq q$, the density function is decreasing and the distribution function is concave. When $p < q$, the density function has a global maximum and the distribution function is S-shaped. *Source: author analysis.*

But $F_T = \sum_{t=0}^{T} f_t$, which implies $f_T = F_T - F_{T-1}$. Equation 2.8 now becomes:

$$F_T = F_{T-1} + [1 - F_{T-1}][p + qF_{T-1}] \tag{2.9}$$

which is the discrete time equivalent of equation 2.4, which is solved using the initial condition $F_0 = 0$.

Final remarks

Bass' model is among the first mathematical models of innovation diffusion. The link with Rogers' theory is direct and explicit, and the model depends on three parameters that are easy to understand: a parameter for innovation, another one for imitation and the market size. Although no graph structure is assumed, the model can be enriched by doing so, and indeed it can be shown that the Bass model is similar to a contagion process over a complete graph (section 2.4). Bass' model is still used today because of its simplicity and the clarity of the mathematical argument.

2.2 Epidemiology: the Diffusion of Pathogens

The spread of a disease among a susceptible population is a diffusion process. Indeed, epidemiology is one of the first fields to have mathematically analysed a diffusion process, and served as a basis for posterior models, such as Bass' (section 2.1.3). The early models did not took network structure under consideration, but it is now an issue known since long (section 2.4). Between 1927 and 1932, William Ogilvy Kermack (1898–1970) and Anderson Gray McKendrick (1876–1943), two Scottish epidemiologists, submitted a series of three papers that established a mathematical theory of the contagion of a disease in a population [29, 30, 31]. Their theory was based on the work by Sir Ronald Ross (1857–1932) (with whom McKendrick had previously worked), who was awarded the Nobel Prize in medicine in 1902 for his work on malaria. Those papers laid the foundations for the mathematical study of epidemics, and are the origin of the well-known SIR and SIS models, along with their variants. As their discussion can provide valuable insight for the problem of the diffusion of information as well, and in particular serve as a basis for the mathematical formulation of the Bass model [16], some aspects of Kermack and McKendrick's papers will be considered below. In particular, the deduction of the SIR model is reproduced.

2.2.1 The SIR model

The SIR model — SIR stands for *susceptible-infected-removed* — was first presented with the publication in 1927 of *A Contribution to the Mathematical Theory of Epidemics* [29]. In words of Kermack and McKendrick, the problem is defined as follows:

> "One (or more) infected person is introduced into a community of individuals, more or less susceptible to the disease in question. The disease spreads from the affected to the unaffected by contact infection. Each infected person runs through the course of his sickness, and finally is removed from the number of those who are sick, by recovery or by death. The chances of recovery or death vary from day to day during the course of the illness. The chances that the affected may convey infection to the unaffected are likewise dependent upon the stage of the sickness. As the epidemic spreads, the number of unaffected members of the community becomes reduced. Since the course of an epidemic is short compared with the life of an individual, the population may be considered as remaining constant, except in as far as it is modified by deaths due to the epidemic disease itself." [29, pp. 700–701]

Furthermore, two assumptions are made: the individuals are all equally susceptible to the disease, and a single infection confers permanent immunity if recovered from [29, p. 701]. In the sequel, the set of discrete time equations governing the course of an epidemic is deduced. Then, the differential equations in continuous time is presented, and finally the special case with constant rates is discussed.

Discrete time equations

In what follows, the quantities are given for a unit area, and therefore must be thought of as densities. Let $v_{t,\theta}$ be the number of individuals who have been sick for θ periods of time at instant t, therefore the total number of the ill at t is $y_t = \sum_{\theta=0}^{t} v_{t,\theta}$ [29, p. 702]. Let v_t be the number of individuals who become sick at t, then $v_{t,0} = v_t \quad \forall t > 0$, with:

$$v_{0,0} = v_0 + y_0 \tag{2.10}$$

since y_0 individuals are initially ill [29, p. 702]. If ψ_θ is the rate of removal (by recovery or by death) of the individuals who have been ill for θ periods, then $v_{t+1,\theta+1} = v_{t,\theta} - \psi_\theta v_{t,\theta} = v_{t,\theta}(1 - \psi_\theta)$ holds. Consequently:

$$v_{t,\theta} = v_{t-\theta,0} B_\theta \tag{2.11}$$

where $B_\theta = \prod_{s=0}^{\theta-1}(1 - \psi_s)$ [29, p. 703].

Now, let ϕ_θ be the infectivity rate after θ periods of illness, with $\phi_0 = 0$ since "an individual is not infective at the moment of infection" [29, p. 703]. The fundamental assumption of the model is that "the chance of an infection is proportional to the number of infected on the one hand, and to the number not yet infected on the other" [29, p. 703]. This leads to:

$$v_t = x_t \sum_{\theta=1}^{t} \phi_\theta v_{t,\theta} \quad \text{[29, p. 703]} \tag{2.12}$$

where x_t is the number of unaffected individuals at t [29, p. 703]. By definition of x_t, the equations $x_t = N - \sum_{s=0}^{t} v_{s,0} = N - \sum_{s=0}^{t} v_s - y_0$ hold, where N is the initial density of

population [29, p. 703]. Moreover, if z_t is the number of those who have been removed either by death or by recovery, then:

$$x_t + y_t + z_t = N \quad \forall t \quad \text{[29, p. 703]} \tag{2.13}$$

Rewriting equation 2.12:

$$\begin{aligned} v_t &= x_t \sum_{\theta=1}^{t} \phi_\theta B_\theta v_{t-\theta,0} \quad \text{by 2.11} \\ &= x_t(\sum_{\theta=1}^{t} \phi_\theta B_\theta v_{t-\theta} + \phi_t B_t y_0) \quad \text{by 2.10} \\ &= x_t(\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0) \quad \text{[29, p. 704]} \end{aligned} \tag{2.14}$$

where $A_\theta = \phi_\theta B_\theta$ [29, p. 704]. Also, as $y_t = \sum_{\theta=0}^{t} v_{t,\theta}$:

$$\begin{aligned} y_t &= \sum_{\theta=0}^{t} B_\theta v_{t-\theta,0} \quad \text{by 2.11} \\ &= \sum_{\theta=0}^{t} B_\theta v_{t-\theta} + B_t y_0 \quad \text{by 2.10 [29, p. 704]} \end{aligned} \tag{2.15}$$

which is an equation for the number of infected individuals at t.

The purpose of the previous calculations is to obtain a set of difference equations for x_t, y_t and z_t, since the point of interest is their dynamics. Since v_t is the number of individuals who become ill at t, then:

$$-v_t = x_{t+1} - x_t \quad \text{[29, p. 704]} \tag{2.16}$$

Therefore equation 2.14 becomes:

$$x_{t+1} - x_t = -x_t(\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0) \quad \text{[29, p. 704]} \tag{2.17}$$

The number of individuals removed at the end of t is by definition $z_{t+1} - z_t$ [29, p. 704], and must be equal to $\sum_{\theta=1}^{t} \psi_\theta v_{t,\theta}$, which is equal by 2.11 and 2.10 to $\sum_{\theta=1}^{t} \psi_\theta B_\theta v_{t-\theta} + \psi_t B_t y_0$ [29, p. 704]. Defining $C_\theta = \psi_\theta B_\theta$ [29] yields:

$$z_{t+1} - z_t = \sum_{\theta=1}^{t} C_\theta v_{t-\theta} + C_t y_0 \quad \text{[29, p. 704]} \tag{2.18}$$

And by 2.13:

$$y_{t+1} - y_t = x_t[\sum_{\theta=1}^{t} A_\theta v_{t-\theta} + A_t y_0] - [\sum_{\theta=1}^{t} C_\theta v_{t-\theta} + C_t y_0] \quad \text{[29, p. 704]} \tag{2.19}$$

Equations 2.16, 2.17, 2.18 and 2.19 describe the dynamics in discrete time of the system.

Continuous time equations

Now, making the time intervals infinitesimal, equation 2.15 becomes:

$$y_t = \int_0^t B_\theta v_{t-\theta} \mathrm{d}\theta + B_t y_0 \quad [29, \text{ p. } 704] \tag{2.20}$$

The equations 2.13, 2.16, 2.17, 2.18 and 2.19 are now written as:

$$x_t + y_t + z_t = N \quad [29, \text{ p. } 704] \tag{2.21}$$

$$v_t = -\frac{\mathrm{d}x_t}{\mathrm{d}t} \quad [29, \text{ p. } 704] \tag{2.22}$$

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -x_t \left[\int_0^t A_\theta v_{t-\theta} \mathrm{d}\theta + A_t y_0 \right] \quad [29, \text{ p. } 704] \tag{2.23}$$

$$\frac{\mathrm{d}y_t}{\mathrm{d}t} = x_t \left[\int_0^t A_\theta v_{t-\theta} \mathrm{d}\theta + A_t y_0 \right] - \left[\int_0^t C_\theta v_{t-\theta} + C_t y_0 \right] \tag{2.24}$$

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = \int_0^t C_\theta v_{t-\theta} + C_t y_0 \quad [29, \text{ p. } 704] \tag{2.25}$$

$$\tag{2.26}$$

where:

$$B_\theta = \mathrm{e}^{-\int_0^\theta \psi(a)\mathrm{d}a} \quad , \quad A_\theta = \phi_\theta B_\theta \quad , \quad C_\theta = \psi_\theta B_\theta \tag{2.27}$$

Equation 2.21 is not independent and can be deduced from 2.23, 2.20 and 2.25.

Constant rates

Supposing that $\phi(\theta) \equiv \kappa$ and $\psi(\theta) \equiv l$ [29, p. 713], equations 2.21, 2.23, 2.24 and 2.25 become:

$$x + y + z = N \quad [29, \text{ p. } 713] \tag{2.28}$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\kappa x y \quad [29, \text{ p. } 713] \tag{2.29}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \kappa x y - l y \quad [29, \text{ p. } 713] \tag{2.30}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = l y \quad [29, \text{ p. } 713] \tag{2.31}$$

where equations 2.29, 2.30 and 2.31 are the ordinary differential equations that compose the SIR model, when time is continuous.

Final remarks

The SIR model, seen above, does not consider population dynamics (births and non-specific deaths), nor does it allow for a recovered individual to become sick again. By allowing partial

immunity for recovered individuals, the SIS model is obtained [30]. Kermack and McKendrick first introduced a global birth rate [30] — composed by immigration and reproduction of unaffected, susceptible and infected individuals —, and then incorporated a non-specific death rate [31].

The above mathematical argument is the first rigorous treatment of the spread of a disease. The key assumption is how the disease spread from infected individuals to the unaffected. By equation 2.3, the number of individuals who fall ill at t is $x_t \sum_{\theta=1}^{t} \phi_\theta v_{t,\theta}$ where x_t is the number of unaffected individuals, ϕ_θ is the infectivity rate at age θ of the illness, and $v_{t,\theta}$ is the number of individuals who have been ill for θ periods at t. This equation suggests that each unaffected individual becomes sick with probability $\sum_{\theta=1}^{t} \phi_\theta v_{t,\theta}$. On the one hand, the sum indicates that each unaffected individual may be infected by an sick one of any age θ. On the other hand, the term $\phi_\theta v_{t,\theta}$ shows that, within category of age θ, any infected individual can spread the disease. In other words, any infected individual can spread the disease to any unaffected individual, for all t. If contagion occurs by contact, the implicit assumption is therefore that all individuals have at least one contact with one another, at any period of time: the implicit assumption is that the contact network is indeed a complete graph.

2.3 The voter model

The voter model[5] is a stochastic model that describes how consensus emerges from a population of individuals [35, p. 93]. It has been studied in the context of stochastic processes, and has been applied to interacting particle systems. For example, the voter model is a first approximation of how an initially disordered ferromagnetic material transitions to an ordered state [35, p. 93]. The dynamical properties of the voter model depend on the size of the system, and of the dimensionality of the arrangement of particles [35, p. 93]. As stated by Redner:

> "In the voter model, individuals are situated at each of the sites of a graph — one for each site. This graph could be a regular lattice in d dimensions, or it could be any type of graph —such as the Erdös-Rényi random graph, or a graph with a broad distribution of degrees. Each voter can be in one of two states that, for this presentation, we label as "Democrat" and "Republican". Mathematically, the state of the voter at $\mathbf{x}$, $s(\mathbf{x})$, can take the values ± 1 only; $s(\mathbf{x}) = +1$ for a Democrat and $s(\mathbf{x}) = -1$ for a Republican." [35, p. 93]

The model then consists of the following rules:

1. "Pick a random voter.

[5]This subsection is based on Chapter 6 of *Fundamental Kinetic Processes* by Redner [35], which can be downloaded at http://physics.bu.edu/∼redner/896

2. The selected voter at $\mathbf{x}$ adopts the state of a randomly-selected neighbor at $\mathbf{y}$. That is, $s(\mathbf{x}) \to s(\mathbf{y})$.

3. Repeat steps 1 & 2 *ad infinitum* or stop when consensus is achieved."[35, p. 93]

For a finite population of voters, the process "eventually achieves consensus in a time that depends on the system size and on the spatial dimension" [35, p. 93]. Below, the probability p that the consensus favors the Democrats will be computed.

The rate at which the voter $\mathbf{x}$ changes of opinion to $s(\mathbf{x})$ is:

$$w(s(\mathbf{x})) = \frac{1}{2}\left(1 - \frac{s(\mathbf{x})}{z}\sum_{\mathbf{y}\, n.n.\mathbf{x}} s(\mathbf{y})\right) \quad \text{[35, p. 94]} \tag{2.32}$$

where "the sum is over the nearest neighbors of site $\mathbf{x}$" [35, p. 94] and "z is the coordination number of the graph" [35, p. 94], which is the same for each site $\mathbf{x}$. Decomposing the neighbors $\mathbf{y}$, according to their agreement or disagreement with $\mathbf{x}$, equation 2.33 becomes:

$$w(s(\mathbf{x})) = \frac{1}{2}\left(1 - \frac{1}{z}\sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})=s(\mathbf{x})}} s(\mathbf{x})s(\mathbf{y}) - \frac{1}{z}\sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})\neq s(\mathbf{x})}} s(\mathbf{x})s(\mathbf{y})\right) \tag{2.33}$$

As $s(\mathbf{y}) = s(\mathbf{x}) \Rightarrow s(\mathbf{y})s(\mathbf{x}) = +1$ and $s(\mathbf{y}) \neq s(\mathbf{x}) \Rightarrow s(\mathbf{y})s(\mathbf{x}) = -1$, then:

$$w(s(\mathbf{x})) = \frac{1}{2}\left(1 + \frac{\sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})\neq s(\mathbf{x})}} 1 - \sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})=s(\mathbf{x})}} 1}{z}\right) \tag{2.34}$$

Let $N_{\mathbf{x}}^{\neq} = \sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})\neq s(\mathbf{x})}} 1$ and $N_{\mathbf{x}}^{=} = \sum_{\substack{\mathbf{y}\, n.n.\mathbf{x} \\ s(\mathbf{y})=s(\mathbf{x})}} 1$, then $N_{\mathbf{x}}^{\neq} + N_{\mathbf{x}}^{=} = z$, and equation 2.34 may be written as:

$$\begin{aligned} w(s(\mathbf{x})) &= \frac{1}{2}\left(1 + \frac{N_{\mathbf{x}}^{\neq} - N_{\mathbf{x}}^{=}}{z}\right) \\ &= \frac{1}{2}\left(1 + \frac{N_{\mathbf{x}}^{\neq} - (z - N_{\mathbf{x}}^{\neq})}{z}\right) \\ &= \frac{1}{2}\left(1 + \frac{2N_{\mathbf{x}}^{\neq} - z}{z}\right) \\ &= \frac{N_{\mathbf{x}}^{\neq}}{z} \end{aligned} \tag{2.35}$$

hence $w(s(\mathbf{x}))$ can be thought of as the fraction of disagreeing neighbors: if all neighbors disagree, then $\mathbf{x}$ changes to state $-s(\mathbf{x})$ with rate 1, otherwise if they all bear the same opinion the change rate is zero.

Now, consider the mean state at $\mathbf{x}$, $S(\mathbf{x}) = \langle s(\mathbf{x})\rangle$, where the mean is computed over all possible configurations: for a general function $f(\{s\})$, $\langle f(\{s\})\rangle = \sum_s f(\{s\})Pr[s]$. As stated by Redner:

"In a small time interval Δt, the state of a given voter changes as follows:

$$s(\mathbf{x}, t+\Delta t) = \begin{cases} s(\mathbf{x},t) & \text{with probability } 1 - w(s(\mathbf{x}))\Delta t, \\ -s(\mathbf{x},t) & \text{with probability } w(s(\mathbf{x}))\Delta t\text{" [35, p. 94]} \end{cases} \tag{2.36}$$

The expected value of $s(\mathbf{x}, t+\Delta t)$ is therefore:

$$\begin{aligned} s(\mathbf{x}, t+\Delta t) &= s(\mathbf{x},t)(1 - w(s(\mathbf{x}))\Delta t) - s(\mathbf{x},t)w(s(\mathbf{x}))\Delta t \\ &= s(\mathbf{x},t) - 2s(\mathbf{x})w(s(\mathbf{x}))\Delta t \end{aligned} \tag{2.37}$$

As $\Delta t \to 0$, equation 2.37 yields:

$$\frac{\mathrm{d}(s(\mathbf{x}))}{\mathrm{d}t} = -2s(\mathbf{x})w(s(\mathbf{x})) \tag{2.38}$$

Therefore:

$$\begin{aligned} \frac{\mathrm{d}S(\mathbf{x})}{\mathrm{d}t} &= \frac{\mathrm{d}\langle s(\mathbf{x})\rangle}{\mathrm{d}t} \\ &= \langle \frac{\mathrm{d}s(\mathbf{x})}{\mathrm{d}t} \rangle \\ &= -2\langle s(\mathbf{x})w(s(\mathbf{x}))\rangle \quad \text{(by 2.38; [35, p. 95])} \end{aligned} \tag{2.39}$$

which gives an equation for the temporal evolution of $S(\mathbf{x})$.

Developing equation 2.39:

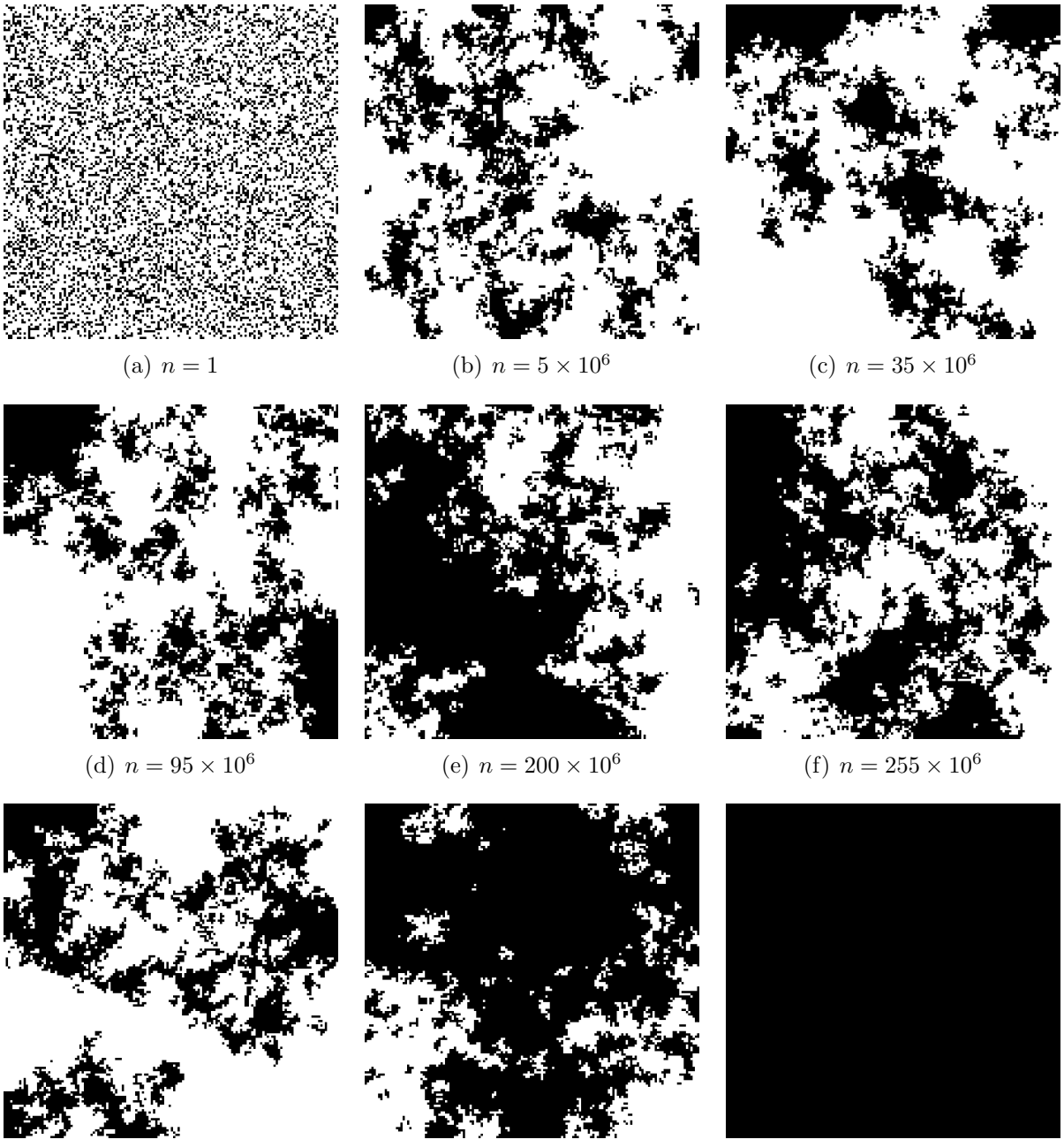
$$\begin{aligned} \frac{\mathrm{d}S(\mathbf{x})}{\mathrm{d}t} &= -2\langle s(\mathbf{x})\frac{1}{2}\big(1 - \frac{s(\mathbf{x})}{z}\sum_{\mathbf{y}\, n.n.\mathbf{x}} s(\mathbf{y})\big)\rangle \quad \text{(by 2.32)} \\ &= -\langle s(\mathbf{x}) - \frac{s(\mathbf{x})^2}{z}\sum_{\mathbf{y}\, n.n.\mathbf{x}} s(\mathbf{y})\rangle \\ &= -S(\mathbf{x}) + \frac{1}{z}\sum_{\mathbf{y}\, n.n.\mathbf{x}} S(\mathbf{y}) \quad (\text{since } s(\mathbf{x})^2 = 1) \\ &= -S(\mathbf{x}) + \frac{1}{z}\sum_{\mathrm{i}} S(\mathbf{x} + \mathbf{e}_\mathrm{i}) \quad \text{[35, p. 95]} \end{aligned} \tag{2.40}$$

where the last equation comes from assuming a d-dimensional lattice, with $\mathbf{e}_\mathrm{i}$ being the unit vectors of the lattice [35, p. 95]. Summing over all sites:

$$\sum_{\mathbf{x}} \frac{\mathrm{d}S(\mathbf{x})}{\mathrm{d}t} = \sum_{\mathbf{x}} S(\mathbf{x}) - \frac{1}{z}\sum_{\mathbf{x}}\sum_{\mathrm{i}} S(\mathbf{x} + \mathbf{e}_\mathrm{i}) \tag{2.41}$$

In the negative term of the right hand side, each site is summed z times since it has z neighbors. Therefore both sums cancel each other:

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{\mathbf{x}} S(\mathbf{x}) = 0 \tag{2.42}$$



(a) $n = 1$ (b) $n = 5 \times 10^6$ (c) $n = 35 \times 10^6$

(d) $n = 95 \times 10^6$ (e) $n = 200 \times 10^6$ (f) $n = 255 \times 10^6$

(g) $n = 375 \times 10^6$ (h) $n = 480 \times 10^6$ (i) $n = 695 \times 10^6$

Figure 2.3: A simulation of the voter model on a square 2D-lattice of size 150×150, with Democrats in black and Republicans in white. The simulation step n is given under each subfigure. Initially, voters are displayed at random, with a probability $\rho = 0.3$ of being Democrat and $1 - \rho = 0.7$ of being Republican (subfigure **(a)**). It turns out that eventually the Democrats won: this event had a probability of $\rho = 0.3$ (subfigure **(i)**). *Source: author analysis.*

Further, dividing by the number of sites N, it results that the mean magnetization $m = \sum_{\mathbf{x}} S(\mathbf{x})/N$ is conserved [35, p. 95]. Since a finite system eventually reaches a consensus [35, p. 95], the probability of its occurrence can be computed as follows.

Suppose that initially, a fraction ρ are Democrats, so that a fraction $1-\rho$ are Republicans. The initial magnetization is then $m_0 = \rho \times 1 + (1-\rho) \times (-1) = 2\rho - 1$ [35, p. 95]. Assuming that with probability E the all-Democrats consensus is reached, then the consensus magnetization is $m_\infty = E \times 1 + (1-E) \times (-1) = 2E - 1$ [35, p. 95]. By conservation of magnetization, equality $2\rho - 1 = 2E - 1$ holds, and so $E = \rho$ [35, p. 95]: the probability of an all-Democrats consensus is equal to the initial fraction of democrat voters (figure 2.3).

Final remarks

Despite its simplicity, the voter model provides a rich framework for the analysis of consensus among complex systems. In particular, the models implemented in this work are variants of the voter model. First, they assume that the users interact with each other through the web forum, thus forming a graph of interaction. Second, when they choose a post for reading, their preferences are modified by the chosen post. There is therefore an incorporation of the chosen post's author's "spin". The models implemented can then be interpreted as voter model processes on general graphs, with a continuous state space (instead of the simpler case ± 1) and with specific spin incorporation rules.

2.4 The importance of the network structure

It has been noticed before than the deduction of both the SIR model and the Bass' model implicitly assume a complete graph as a the medium of propagation (see sections 2.1.3 and 2.2.1). In the present section, a stochastic process equivalent to the Bass model is deduced and tested on random graphs of varying densities. The results obtained show that when the graph is complete, the adjustment of the stochastic process with Bass' theoretical equation is very good. When the graph is not, a slower adoption curve is obtained only if the parameter of contagion is relevant. When the latter is low, no effect is noticeable, as the adoption proceeds exclusively by innovation, and therefore the graph plays no role.

Recalling the SIR model of Kermack and McKendrick, in particular equation 2.12:

$$v_t = x_t \sum_{\theta=1}^{t} \phi_\theta v_{t,\theta} \tag{2.43}$$

where v_t is the number of individuals per unit area that become ill at t, x_t is the number of unaffected individuals per unit area, ϕ_θ is the infectivity rate at age θ and $v_{t,\theta}$ is the number

of infected individuals that have been ill for θ periods. Supposing that the infectivity rate is constant $\phi_\theta \equiv \kappa$, and that individuals just infected are contagious, equation 2.43 becomes:

$$v_t = x_t \kappa y_t \tag{2.44}$$

where $y_t = \sum_{\theta=0}^{t} v_{t,\theta}$ is the total number of infected individuals per unit area at t. By equation 2.16, $v_t = -(x_{t+1} - x_t)$. But, assuming a zero removal rate (an infected individual remains contagious permanently), then the number of removed individuals per unit area is constantly zero: $z_t \equiv 0$. Therefore, from equation 2.13 it can be deduced that $x_t = N - y_t$ and $y_{t+1} - y_t = -(x_{t+1} - x_t)$, so equation 2.44 may be written as:

$$\begin{aligned} y_{t+1} - y_t &= (N - y_t)\kappa y_t \\ \Leftrightarrow y_{t+1} &= y_t + (N - y_t)\kappa y_t \end{aligned} \tag{2.45}$$

The term $(N - y_t)\kappa y_t$ is interesting, in the sense that it suggests that each unaffected individual — there are $(N - y_t)$ of them — gets ill with probability κy_t. So, each infected individual has an equal chance of infecting an unaffected individual. Kermack and McKendrick assumed no underlying structure in the contagion process, but in case they did, it would have been a complete graph: each individual is linked with all the rest. Now it is known that diseases spread through contact networks, and that the complete graph is not a realistic assumption.

With regard to the Bass model, the discrete dynamical equation may be written as:

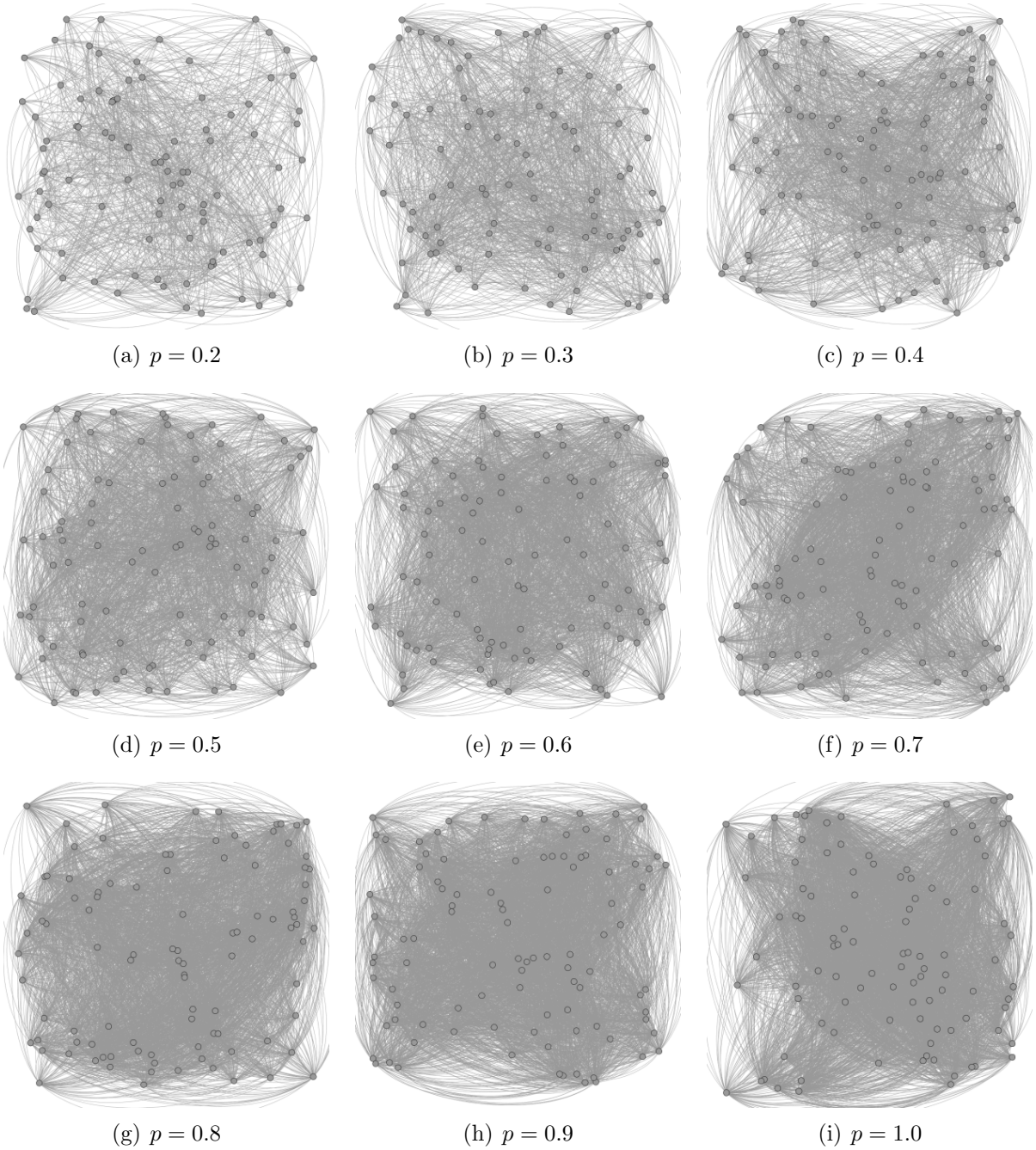
$$\begin{aligned} F_{t+1} &= F_t + [1 - F_t][p + qF_t] \quad \text{(by 2.9)} \\ \Leftrightarrow mF_{t+1} &= mF_t + [m - mF_t][p + \frac{q}{m} mF_t] \\ \Leftrightarrow mF_{t+1} &= mF_t + p[m - mF_t] + [m - mF_t]\frac{q}{m} mF_t \end{aligned} \tag{2.46}$$

where m is the total number of adopters, F_t is the probability of the innovation having been adopted before t, p is the coefficient of innovation and q is the coefficient of imitation. As $Y_t = mF_t$ is the number of those who have adopted the innovation before t, equation 2.46 becomes:

$$Y_{t+1} = Y_t + p[m - Y_t] + [m - Y_t]\frac{q}{m} Y_t \tag{2.47}$$

Putting aside the innovation term $p[m - Y_t]$, equation 2.47 is the exact equivalent of equation 2.45. This is no discovery since Bass was inspired by the mathematical theory of epidemics [16, p. 215]. But it must be noted again that no underlying structure is considered, and that the complete graph is the best analogue.

The equivalence with the complete graph will be proven empirically in the following. Consider a stochastic process over a graph, that simulates the spread of an epidemic over a graph with m nodes. All nodes are initially unaffected, and become ill spontaneously with probability p (innovation process). Also, an infected node may transfer the disease to each of his neighbors with probability q, each time. This process intends to replicate the dynamics of the discrete Bass model (equation 2.47) over a graph, and to compare the cumulative adoption curves. For this purpose, the process is run for distinct values of p and q, over a



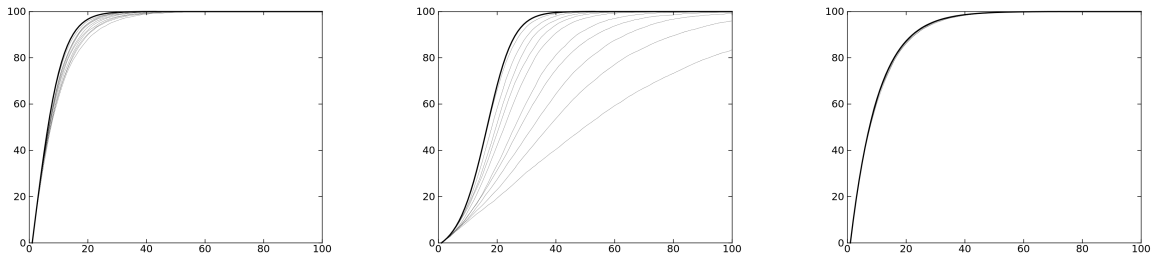
(a) $p = 0.2$ (b) $p = 0.3$ (c) $p = 0.4$

(d) $p = 0.5$ (e) $p = 0.6$ (f) $p = 0.7$

(g) $p = 0.8$ (h) $p = 0.9$ (i) $p = 1.0$

Figure 2.4: Erdös-Rényi graphs, with edge probability p_E ranging from 0.1 (subfigure **(i)**) to 1.0 (subfigure **(i)**). *Source: author analysis.*

set of Erdös-Rényi graphs. An Erdös-Rényi graph is generated at random, where each edge exists with a fixed probability p_E ([25]; see figure 2.4). As can be seen in figure 2.5, as the probability p_E increases, the adoption curve tends towards the Bass curve, with a very good fit obtained when $p_E = 1$ — the graph is complete. It can be therefore said that the models of Kermack and McKendrick on one hand, and of Bass on the other hand, assume implicitly a complete graph structure.



(a) $m = 100, p = 0.1, q = 0.1$ (b) $m = 100, p = 0.01, q = 0.2$ (c) $m = 100, p = 0.1, q = 0.005$

Figure 2.5: Comparison between Bass adoption curves and the stochastic process over Erdös-Rényi graphs, with the edge probability p_E ranging from 0.1 to 1. Subfigures **(a)** and **(b)** correspond to the cases of figure 2.2. Subfigure **(b)** shows than when the imitation coefficient q dominates, the rate of adoption is very dependent upon the density of the graph: the more the graph is dense, the faster the adoption spreads. When p and q are equal, the network structure has a lesser impact as shown in subfigure **(a)**. If q is very small, then the adoption spreads almost exclusively through the innovation effect, and the network effect becomes invisible (subfigure **(c)**).

Chapter 3

Decision Making Models

The models implemented in this work simulate text choices made by the OSN users. Indeed, the web forum is simulated as a list of threads, and each thread is modeled in turn as list of posts (chapter 5). The choice of the thread must consider the text content of its first post (text compatibility) and the publication date of its last post (recency). On the other hand, the choice of the post must consider in addition the social image of the poster (image compatibility). Therefore, a decision making model is needed in order to implement those decision mechanisms. The main decision making model used in this work is the leaky competing accumulator [44], first described below. Then, in order to provide a reasonable benchmark, the classic logit model is discussed. While the leaky competing accumulator takes the form of a set of stochastic equations that must be simulated in order to determine the choice, the logit model provides directly the choices' probabilities which are derived from a set of axioms.

3.1 Perceptual choice: The leaky, competing accumulator model

In chapter 1, section 1.1.3, a brief account of some developments in the field of perceptual choice was made, developments which lead eventually to the model that is of interest in this work, the *leaky competing accumulator* model. The previous account will therefore be assumed as solid ground for the following discussion, and focus will be put on this model. Perceptual choice deals with the problem of understanding how the brain processes the information proceeding from the senses, and decides between alternatives accordingly. In particular, two variables that are to be explained are which decision is taken, and the duration of the decision process. Typical questions that arise in perceptual choice are to determine the influence of the number of alternatives in both the duration of the decision process and its accuracy, or to study the relation between the duration of the decision process and the accuracy achieved. Therefore, perceptual choice is of interest for this work, as an important component of the models that will be presented and evaluated in the next chapters depend on the decisions made by the users of an online social network.

The leaking competing accumulator was first proposed by Usher and McClelland in 2001 [44], and consists of a set of stochastic equations governing the evidence accumulated in favor of N alternatives. In what follows, the set of equations will be deduced and explained, and some aspects of its dynamics will be analyzed. The deduction of the equation is based on a publication by Usher and McClelland [44], and the analysis of the model's dynamics is inspired by the work of Bogacz et al. [18]. The analysis of the LCA dynamic will play an important role in the settings of the main model implemented in this work (chapter 5, section 5.2.1).

3.1.1 The equations governing the LCA model

According to the discussion by Usher and McClelland, six principles of information processing by populations of neurons can be formulated, and classified depending on if they are *early work* principles, *main* principles. or *additional* principles:

early work principle 1(EW1): *accumulation over time*; "[models] treat information processing as a gradual process, based on the accumulation of information over time" [44, p. 550] (discussed at length in pp. 550–552)

early work principle 2(EW2): *variability*; "[models] treat the process as stochastic or intrinsically variable, so that the information accumulated within each small time interval is subject to random fluctuations" [44, p. 550] (discussed at length in pp. 550–552 and p.555)

main principle 1(M1): *leakage/decay*; "information accumulation is subject to leakage or decay" [44, p. 550] (discussed at length in p. 555)

main principle 2(M2): *lateral inhibition*; "representation of the alternative outcomes of the decision process compete with each other, through a process of lateral inhibition" [44, p. 550] (discussed at length in pp. 555–556)

additional principle 1(A1): *recurrent excitation*; "modelers working in a neuroscience framework have suggested that recurrent excitation may play a prominent role in maintaining activity in neural populations" [44, p. 555]

additional principle 2(A2): *nonlinearity*; "many computations thought to be essential for perception , cognition, and action cannot be carried out without at least one layer of nonlinear computation" [44, p. 553]

Keeping the above principles in mind, the case of N alternatives is considered, and the set of equations of the LCA model is deduced. For this purpose, two classes of cognitive units are

at work: accumulator units, and input units[1]. Each alternative $\mathrm{i} \in \{1, \ldots, N\}$ is represented by an accumulator unit, whose activation is x_i. The starting point is the following equation [44, p. 558]:

$$\mathrm{d}x_\mathrm{i} = [I_\mathrm{i} - \lambda x_\mathrm{i}]\frac{\mathrm{d}t}{\tau} + \chi_\mathrm{i}\sqrt{\frac{\mathrm{d}t}{\tau}} \tag{3.1}$$

Concerning the explanation of the terms involved, the authors state: "τ is a time scale chosen for convenience, and χ_i is a Gaussian noise term with zero mean and variance σ^2. This equation implies that within a time interval $\mathrm{d}t/\tau$, the change in the activation of an accumulator unit, $\mathrm{d}x_\mathrm{i}$, is driven by input from other units, I_i, with a characteristic decay rate λ, which reflects leakage of the activation. The noise term scales with the square root of $\mathrm{d}t/\tau$, because the variance of uncorrelated stochastic random variables is additive, leading to the square-root behavior for the standard deviation." [44, p. 558].

In equation 3.1, the principles **EW1**, **EW2** and **M1** are present. In effect, the accumulation-over-time principle is reflected in the increment (or decrement) of $\mathrm{d}x_\mathrm{i}$ in the activation level for alternative i, over a time interval of $\mathrm{d}t$. On the other hand, $\chi_\mathrm{i}\sqrt{\frac{\mathrm{d}t}{\tau}}$ is normally distributed with zero mean and variance equal to $\sigma^2\mathrm{d}t/\tau$, thus representing the variability principle. Finally, the leakage principle can be appreciated in the term $-\lambda x_\mathrm{i}$, which becomes increasingly negative as x_i gets bigger. Yet, the principles **D2**, **A1** and **A2** are still missing in equation 3.1, so that further manipulation is needed.

At this point of the analysis, the key resides in the input term I_i. According to the words of Usher and McClelland: "the input I_i can be decomposed into three distinct components: an external source, I_i^{ext}, a recurrent excitatory source, I_i^{rec}, coming from the unit back to itself, and lateral inhibition between accumulator units" [44, p. 558]. If the lateral inhibition is denoted by LI_i, then the previous phrase suggests:

$$I_\mathrm{i} = I_\mathrm{i}^{ext} + I_\mathrm{i}^{rec} - LI_\mathrm{i} \tag{3.2}$$

Furthermore, the external source is rewritten as ρ_i; the recurrent excitation is defined as $I_\mathrm{i}^{rec} = \alpha f(x_\mathrm{i})$, where α is a scaling factor and $f(\cdot)$ is the *threshold-linear* function[2]; and finally, the total lateral inhibition exerted on the accumulator unit i is $LI_\mathrm{i} = -\sum_{j\neq\mathrm{i}}\beta_{j\mathrm{i}}f(x_j) = -\beta\sum_{j\neq\mathrm{i}}f(x_j)$, as the simplifying assumption $\beta_{\mathrm{i}j} = \beta$ is adopted. Thus, equation 3.2 becomes:

$$I_\mathrm{i} = \rho_\mathrm{i} + \alpha f(x_\mathrm{i}) - \beta\sum_{j\neq\mathrm{i}} f(x_j) \tag{3.3}$$

[1]Concerning cognitive units and activation, Usher and McClelland state: "We adopt the dominant approach taken in computational neuroscience today, which follows the Hebbian perspective (Hebb, 1949). In this approach, each cognitive unit is represented by a pattern of activation over a group of neurons, or *cell population*, and the activation of the cognitive unit is represented by the mean firing rate of the neurons in the population" [44, p. 554].

[2]The threshold-linear function is defined as $f(x) = x$ if $x \geq 0$, $f(x) = 0$ otherwise

The stochastic equations 3.1 are now rewritten as:

$$\mathrm{d}x_\mathrm{i} = [\rho_\mathrm{i} - \lambda x_\mathrm{i} + \alpha f(x_\mathrm{i}) - \beta \sum_{j \neq \mathrm{i}} f(x_j)]\frac{\mathrm{d}t}{\tau} + \chi_\mathrm{i}\sqrt{\frac{\mathrm{d}t}{\tau}} \tag{3.4}$$

Approximating $f(x)$ by x when $x \geq 0$, or truncating to zero if $x < 0$, the benefits of a set of linear equations are obtained, without the problem of unphysiological negative activation levels x_i [44, p. 558]. The set of LCA equations are then [44, p. 559]:

$$\begin{cases} \mathrm{d}x_\mathrm{i} = [\rho_\mathrm{i} - (\lambda - \alpha)x_\mathrm{i} - \beta \sum_{j \neq \mathrm{i}} x_j]\frac{\mathrm{d}t}{\tau} + \chi_\mathrm{i}\sqrt{\frac{\mathrm{d}t}{\tau}} \\ x_\mathrm{i} \rightarrow \max(x_\mathrm{i}, 0) \end{cases} \tag{3.5}$$

where $\lambda - \alpha \geq 0$ is the *net leakage*. It is finally noticed that the six principles are now included, as shown in table 3.1.

information processing principle	representing term in equation 3.5
accumulation over time (EW1)	$\mathrm{d}x_\mathrm{i}$ and $\mathrm{d}t/\tau$
variability (EW2)	$\chi_\mathrm{i} \sim N(0, \sigma^2)$
leakage/decay (M1)	$-\lambda x_\mathrm{i}$
lateral inhibition (M2)	$-\beta \sum_{j \neq \mathrm{i}} x_j$
recurrent excitation (A1)	αx_i
nonlinearity (A2)	$x_\mathrm{i} \rightarrow \max(x_\mathrm{i}, 0)$

Table 3.1: Information processing principles within the set of LCA equations 3.5. *Source: Usher and McClelland [44].*

3.1.2 An analysis of the LCA dynamics

Now, the analysis of the LCA model by Bogacz et al. will be considered [18]. They write the set on nonlinear equations 3.4 for N alternatives as [18, p. 1659]:

$$\mathrm{d}y_\mathrm{i} = (-ky_\mathrm{i} - w \sum_{\substack{j=1 \\ j \neq \mathrm{i}}}^{N} f(y_j) + I_\mathrm{i})\mathrm{d}t + c_\mathrm{i}\mathrm{d}W_\mathrm{i} \tag{3.6}$$

In the above equation, y_i are the activation levels, k is the *decay parameter*, w is the *inhibition parameter*, $f(.)$ is a nonlinear function (which may be *threshold linear*, *piecewise linear* or *sigmoidal*; [18, p. 1659]), I_i are the mean inputs, c_i are the amplitudes of the perturbations and $\mathrm{d}W_\mathrm{i}$ are independent Wiener processes [18, p. 1656]. Comparing with equation 3.4, it can noted that the recurrent excitation is assumed to be linear, and then k is equal to the net leakage $\lambda - \alpha$. On the other hand, $\mathrm{d}t/\tau$ has been replaced by $\mathrm{d}t$ and $c_\mathrm{i} = \sigma$, so that $\mathrm{d}W_\mathrm{i} \sim N(0, \mathrm{d}t)$ and $c_\mathrm{i}\mathrm{d}W_\mathrm{i} = \chi_\mathrm{i}\sqrt{\mathrm{d}t/\tau}$. Finally, ρ_i are rewritten as I_i, and x_i as y_i. Assuming the threshold linear function, the approximation of equation 3.6 (and therefore the equivalent of equation 3.5) yields:

$$\begin{cases} \mathrm{d}y_\mathrm{i} = (-ky_\mathrm{i} - w \sum_{\substack{j=1 \\ j \neq \mathrm{i}}}^{N} y_j + I_\mathrm{i})\mathrm{d}t + c_\mathrm{i}\mathrm{d}W_\mathrm{i} \\ y_\mathrm{i} \rightarrow max(y_\mathrm{i}, 0) \end{cases} \tag{3.7}$$

referred to by Bogacz et al. as the *bounded* LCA model [18, p. 1660].

Before analysing the dynamics implied by equations 3.7, the performance of the linear LCA model is briefly discussed. In the case $N = 2$, the linear LCA model is optimal when $k = w$ holds[3], and both are high [18, pp. 1658–1659]. Indeed, "when the linear LCA model of choice between two alternatives is balanced and both inhibition and decay are high, the model approximates the optimal SPRT [sequential probability ratio test] and makes the fastest decisions for fixed ERs [error rates]" [18, p. 1659]. Furthermore, "the SPRT is optimal in the following sense: among all possible procedures for solving this choice problem giving certain ER, it minimizes the average DT [decision time]" [18, p. 1658–1659]. For the general case of N alternatives, the performance is also maximized in the balanced case, although it is not optimal anymore [18, p. 1659].

In the case of the bounded linear LCA model, evidence shows that it tends to outperform the unbounded linear LCA model when the number of alternatives grows [18, p. 1661]. Moreover, in the case of $N = 2$ alternatives, the performance is maximized when the model balanced, and both k and w are large [18, p. 1660]. In the general case of N alternatives, it is not clear whether the previous conclusion still holds, but in the scope of this work, k and w will be thought of as equal (indeed, the values $k = w = 10$ will be used in the models).

Now, the dynamics of the linear LCA model are discussed. It must be noted that, in the scope this work, the bounded linear LCA model was implemented. However, the discussion is still valid as the focus of attention is put on the positive ranges of the activation levels x_{i}, where the linear and the bounded linear models are identical. On the other hand, the analysis below has important consequences in the empirical settings of the parameters of the LCA model (chapter 5, section 5.2.1).

In the case of $N = 2$ alternatives, the dynamics have been already analysed by Bogacz et al. [18, pp. 1657–1658]. Their analysis is reproduced here, as a useful introduction to the general N case that is discussed thereafter. The two dimensional LCA model is:

$$\begin{cases} \mathrm{d}y_1 = (-ky_1 - wy_2 + I_1)\mathrm{d}t + c_1\mathrm{d}W_1 \\ \mathrm{d}y_2 = (-ky_2 - wy_1 + I_2)\mathrm{d}t + c_2\mathrm{d}W_2 \\ y_1(0) = y_2(0) = 0 \end{cases} \tag{3.8}$$

Or in matrix form:

$$\begin{bmatrix} \mathrm{d}y_1 \\ \mathrm{d}y_2 \end{bmatrix} = \left(- \begin{bmatrix} k & w \\ w & k \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \right)\mathrm{d}t + \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \begin{bmatrix} \mathrm{d}W_1 \\ \mathrm{d}W_2 \end{bmatrix} \quad y(0) = 0 \tag{3.9}$$

A linear transformation is applied in order to uncouple the previous system of equations.

[3]When the decay and inhibition parameters are equal (i.e., $k = w$), the LCA model is called *balanced* [18, p. 1658]

The rotation matrix with angle θ counter clockwise in the two dimensional plane is:

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{3.10}$$

Applying a rotation of the axis by 45° clockwise, the new coordinates are:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos(-45°) & \sin(-45°) \\ -\sin(-45°) & \cos(-45°) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{3.11}$$

$$\Leftrightarrow \begin{cases} x_1 = \frac{y_1 - y_2}{\sqrt{2}} \\ x_2 = \frac{y_1 + y_2}{\sqrt{2}} \end{cases} \tag{3.12}$$

Differentiating the system 3.11 yields:

$$\begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \mathrm{d}y_1 \\ \mathrm{d}y_2 \end{bmatrix} \tag{3.13}$$

Using equation 3.9, and reordering the terms, leads to:

$$\begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{bmatrix} = \left(- \begin{bmatrix} (k-w) & 0 \\ 0 & k+w \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \right) \mathrm{d}t + \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \mathrm{d}W_1 \\ c_2 \mathrm{d}W_2 \end{bmatrix} \tag{3.14}$$

Equation 3.14 define the dynamics in the transformed coordinates system. These dynamics are analysed below along each transformed axis, in the balanced case, ignoring the random perturbations.

Axis x_1 Since in the balanced case $k = w$ holds, then by 3.14 $\frac{\mathrm{d}x_1}{\mathrm{d}t} \equiv \frac{I_1 - I_2}{\sqrt{2}}$: thus, the sign of $\frac{\mathrm{d}x_1}{\mathrm{d}t}$ is equal to the sign of the difference $I_1 - I_2$. Therefore, if $I_1 > I_2$, then the deterministic dynamics tend increase the value of x_1, until the threshold is reached and the alternative 1 is chosen. If $I_1 < I_2$, then the alternative 2 tends to be chosen (figure 3.1). Thus, the choice between alternatives 1 and 2 occur along the axis x_1.

Axis x_2 The dynamics along the x_2 axis are attracted to an equilibrium characterized by:

$$\left. \frac{\mathrm{d}x_2}{\mathrm{d}t} \right|_{x_2^*} = 0 \tag{3.15}$$

$$\Leftrightarrow -(k+w)x_2^* + \frac{I_1 + I_2}{\sqrt{2}} = 0 \qquad \text{(by 3.14)} \tag{3.16}$$

$$\Leftrightarrow x_2^* = \frac{I_1 + I_2}{\sqrt{2}(k+w)} > 0 \tag{3.17}$$

Moreover, this equilibrium is stable as by 3.14 $\left. \frac{\mathrm{d}x_2^2}{\mathrm{d}t^2} \right|_{x_2^*} = -(k+w) < 0$. The line of equation $x_2 = x_2^*$ is therefore an attractor of the dynamics in the (x_1, x_2) plane, and by 3.12 its equation

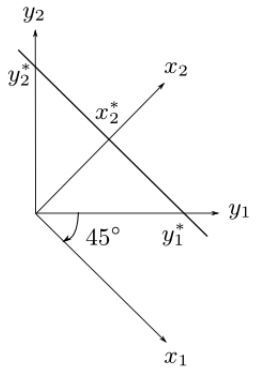


Figure 3.1: By applying a rotation of 45° clockwise on the original coordinates (y_1, y_2), the dynamics of the LCA model are uncoupled: the evolution of x_1 and x_2 are independent. In the $(x1, x2)$ coordinates, the dynamics occur along $x_2 = x_2^*$, where $x_2^* = (I_1+I_2)/(\sqrt{2}(k+w))$ (equation 3.17). The previous line intersects with the axis y_1 and y_2 at $y^* = (I_1+I_2)/(k+w)$ (equation 3.18).

in the (y_1, y_2) plane is $y_1/\sqrt{2} + y_2/\sqrt{2} = x_2^*$, which intersects with the axis (y_1, y_2) at (figure 3.1):

$$y_1^* = y_2^* = y^* = \sqrt{2}x_2^* = \frac{I_1 + I_2}{k + w} \qquad \text{(by 3.17)} \tag{3.18}$$

By equation 3.14, note that in two-dimensional case, ignoring the inputs vector I and the stochastic noise, the following equations were obtained:

$$\begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{bmatrix} = -\begin{bmatrix} k - w & 0 \\ 0 & k + w \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mathrm{d}t \tag{3.19}$$

Let:

$$R = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, D = \begin{bmatrix} k - w & 0 \\ 0 & k + w \end{bmatrix} \tag{3.20}$$

Then, equation 3.19 is equivalent to:

$$\mathrm{d}x = -Dx\mathrm{d}t \qquad \text{(by 3.20)} \tag{3.21}$$

$$\Leftrightarrow R\mathrm{d}y = -DRy\mathrm{d}t \qquad \text{(by 3.20 and 3.11)} \tag{3.22}$$

$$\Leftrightarrow \mathrm{d}y = -R^{-1}DRy\mathrm{d}t \tag{3.23}$$

But equation 3.9 can be rewritten as follows, ignoring the inputs and the stochastic noises:

$$\mathrm{d}y = -\Omega y\mathrm{d}t \tag{3.24}$$

where Ω is a 2×2 matrix such that:

$$\Omega_{ij} = \begin{cases} k & \text{if i} = j \\ w & \text{otherwise} \end{cases} \tag{3.25}$$

Thus, to uncouple the equations 3.9 is equivalent to diagonalize the Ω matrix. It can be shown [38] that in the general case of N alternatives — Ω is then an $N \times N$ matrix —, Ω can be diagonalized as $\Omega = RDR$, where R and D are defined by:

$$R_{\mathrm{i}j} = \frac{1}{\sqrt{n}}(\cos(\frac{2\pi \mathrm{i} j}{n}) + \sin(\frac{2\pi \mathrm{i} j}{n})) \quad 1 \leq \mathrm{i}, j \leq n \tag{3.26}$$

$$D_{\mathrm{i}j} = \begin{cases} k - w & \text{if } \mathrm{i} = j < n \\ k + (n-1)w & \text{if } \mathrm{i} = j = n \\ 0 & \text{otherwise} \end{cases} \tag{3.27}$$

Considering the I inputs and stochastic perturbations $\mathrm{d}W$, the LCA model in the transformed coordinates is:

$$\mathrm{d}x = (-Dx + RI)\mathrm{d}t + Rc\mathrm{d}W \tag{3.28}$$

In the balanced case ($k = w$), and ignoring the stochastic components, the following holds by equations 3.26, 3.27 and 3.28:

$$\frac{\mathrm{d}x_{\mathrm{i}}}{\mathrm{d}t} \equiv < R_{\mathrm{i}\bullet}, I > \quad \mathrm{i} < n \tag{3.29}$$

$$x_n^* = \frac{< R_{n\bullet}, I >}{k + (n-1)w} \quad \text{solves} \quad \frac{\mathrm{d}x_n}{\mathrm{d}t} = 0 \tag{3.30}$$

But $\forall j = 1, \ldots, n \quad R_{nj} = \frac{1}{\sqrt{n}}(\cos(\frac{2\pi n j}{n}) + \sin(\frac{2\pi n j}{n})) = \frac{1}{\sqrt{n}}(\cos(2\pi j) + \sin(2\pi j)) = \frac{1}{\sqrt{n}}$, thus:

$$x_n^* = \frac{\sum_{\mathrm{i}=1}^n I_{\mathrm{i}}}{\sqrt{n}(k + (n-1)w)} \tag{3.31}$$

which is equivalent to the plane (in the $(y_1, \ldots, y_n)$ coordinates):

$$\sum_{\mathrm{i}=1}^n \frac{y_{\mathrm{i}}}{\sqrt{n}} = \frac{\sum_{\mathrm{i}=1}^n I_{\mathrm{i}}}{\sqrt{n}(k + (n-1)w)} \tag{3.32}$$

$$\Leftrightarrow \sum_{\mathrm{i}=1}^n y_{\mathrm{i}} = \frac{\sum_{\mathrm{i}=1}^n I_{\mathrm{i}}}{k + (n-1)w} \tag{3.33}$$

since $x = Ry$, and in particular $x_n =< R_{n\bullet}, y >$. The previous is therefore the attractor hyper-plane of the dynamics of the LCA model, which intersects with any axis y_{i} at $y^* = \frac{\sum_{\mathrm{i}=1}^n I_{\mathrm{i}}}{k + (n-1)w}$.

3.2 The logit model

The multinomial logit model[4] allows to compute the probability to choose an alternative, given the attributes of all the available attributes [26, p. 206]. In the sequel, the fundamental assumptions of the model will be presented, the choice probabilities deduced and the

[4]This subsection is based on *A Logit Model of Brand Choice Calibrated on Scanner Data*, by Guadagni and Little [26]

linear utility shown.

Supposing that the individual i deals with an alternatives set S_i, then the utility of alternative $k \in S_\mathrm{i}$ is [26, p. 207]:

$$u_k = v_k + \varepsilon_k \tag{3.34}$$

where v_k is the deterministic component of u_k, and ε_k is the random component [26, p. 207]. Then, "confronted by the set of alternatives, individual i chooses the one with the highest utility on the occasion" [26, p. 207]. Therefore, the probability of choosing alternative k is:

$$p_k = P[u_k \geq u_j, j \in S_\mathrm{i}] \tag{3.35}$$

Finally, the random components ε_k are independent and identically distributed, with a Gumbel type II extreme value distribution [26, p. 207]:

$$P[\varepsilon_k \leq \varepsilon] = \mathrm{e}^{\mathrm{e}^{-\varepsilon}} \quad , \quad -\infty < \varepsilon < \infty \tag{3.36}$$

Equations 3.34, 3.35 and 3.36 are the axiomatic foundations of the logit model. It can be shown that assuming these three equations, the choice probabilities are written:

$$p_k = \frac{\mathrm{e}^{v_k}}{\sum_{j \in S_\mathrm{i}} \mathrm{e}^{v_\mathrm{i}}} \quad \text{[26, p. 207]} \tag{3.37}$$

where p_k is "S-shaped in v_k when other v_j are held constant" [26, p. 208].

A linear form for the utility is considered in this work, where preferences v_k are a function of the attributes of each alternative [26, p. 209]. Let T_k be the set of attributes unique to alternative k, and T_C the set of attributes common to all alternatives [26, p. 209]. Defining x_{jk}^i as the "observed value of attribute j of alternative k for customer i" [26, p. 209], and b_{jk} as the "the utility weight of attribute j of alternative k" [26, p. 209], then v_k^i may be written as:

$$v_k^\mathrm{i} = \sum_{j \in T_k} b_{jk} x_{jk}^\mathrm{i} + \sum_{j \in T_C} b_j x_{jk}^\mathrm{i} \quad \text{[26, p. 209]} \tag{3.38}$$

Making $T = T_k \cup T_C$, equation 3.38 becomes:

$$v_k^\mathrm{i} = \sum_{j \in T} b_{jk} x_{jk}^\mathrm{i} \quad \text{[26, p. 209]} \tag{3.39}$$

Therefore, in equation 3.37:

$$\mathrm{e}^{v_k} = \prod_{j \in T} \mathrm{e}^{b_{jk}} x_{jk} \quad \text{[26, p. 209]} \tag{3.40}$$

so "the model is, in an important sense, more multiplicative than additive" [26, p. 209].

Chapter 4

Information retrieval for Preference Extraction

The model of information diffusion discussed in this work includes a web forum simulator. Users navigate through the forum structure and interact with its contents. They can browse threads, read posts and publish new messages. A model of text representation is therefore needed, as the simulator must generate new contents. The automatic generation of word sequences that are coherent semantically has not been envisaged, but it is probably a difficult task. Therefore, an alternative approach has been chosen, where posts are modeled as numeric vectors, following the tradition of information retrieval (IR). In particular, an advanced probabilistic model, the latent Dirichlet allocation (LDA, [17]), was used. The use of vectors enables a concise representation of posts generated prior to the simulation, and allows the creation of new posts through vector operations.

IR originates in the historical context of overcrowding libraries and nascent computing power in the 1970s, but rose to prominence with the advent of the World Wide Web. To give a glimpse of the magnitudes involved, in 2005 the number of web pages indexed by major search engines was estimated to be as high as 11.5 billion [27][1]. The current number is probably much higher, so a manual search is actually infeasible. Moreover, web text data are highly unstructured, in contrast with the clear semantics and data format of a relational database. Hence automated searching is a necessity in such conditions, in order to keep the web searchable, and therefore useful. Before discussing some basic IR background and the LDA probabilistic model, a motivating example is presented below.[2]

[1]Yet, the problem of ever growing information volume is not a new one. As Salton and McGill noted in 1983: "by the year 1800, the amount of scientific publication was already doubling every 50 years. More recently with the impressive growth of science and technology, the rate of increase of available knowledge has vastly accelerated. Between 1800 and 1966, the number of scientific journals has increased from 100 to over 100,000. At the present time, no upper limit is apparent in the rate of increase of available information items." [40, p. 3]. Tarde's insight on the natural tendency of innovations towards a geometric progression (chapter 2, section 2.1.1) seems to prove true in this case: not only human population has been growing geometrically, but also the volume of information.

[2]The discussion is inspired in Chapter 1 of Salton and McGill [40], and Chapter 1 of Baeza-Yates and Ribeiro-Neto [15].

How does a user find a book suited for his needs in a library? If he already knows what specific book he is looking for, he may go a to library computer, enter the book title, get its physical location and then go for it. Although the user is only aware of the *physical organization* of the library shelves, the *logical organization* of the books database operates beneath the system's response. The user, motivated by an *information need*, prompts a *query*, which is processed by an *information retrieval (IR) system*. The IR system hopefully returns the most *relevant* information item. In this case, the query consists of the exact title of the book, so the information retrieval task is quite straightforward.

However, the query may consist of the incomplete title, or the book's author, or a set of keywords that represents well the searched book, from the user's perspective. The information retrieval task is much more difficult in this case, particularly if a set of keywords is prompted. A first solution would be to match the keywords with the full text of each book in the collection, with the corresponding prohibitive computational costs. Another solution would be to match the keywords with a set of words that *index* each book. A similarity measure would then be computed between the prompted keywords and each book's *index terms*. Finally, the results would be ordered according to their similarity with the query, and presented to the user.

Two difficulties arise here. First, the matching similarity and ranking process must meet the user's information need, which is not evident. Second, the system requires the assignment of index terms for each book of the collection, which is called *indexing*. Manual indexing can be performed by trained experts, but automatic indexing is mandatory for large collections. Successfully meeting the user information needs and smartly indexing a collection's items are therefore two central challenges of an IR system. In this chapter, a brief review of classic IR models is performed, among three main categories: boolean models, vector models and probabilistic models. Then, the LDA model is described.

4.1 Some background on IR

Information retrieval[3] is defined by Salton and McGill as follows: "information retrieval is concerned with the representation, storage, organization, and accessing of information items" [40, p. 1]. The previous definition may have since become incomplete because of richer WWW capabilities [15, p. 2], but still holds for the classical IR core. A defining characteristic of information retrieval is that it searches in text documents, which are unstructured data [40, p. xi]. This differentiates an IR system from, say, a database management system (DBMS; [40, p. 8]).

[3]This section is based on Chapter 2 of *Modern Information Retrieval* by Baeza-Yates and Ribeiro-Neto [15].

A *term* is defined as a word, and the set of all possible terms is called the *vocabulary*. A *document* is a collection of terms[4] that constitute a single unit: it may be an article, a post in a web forum, a book, and so on, depending on the situation. Finally, a *corpus* is a collection of documents (figure 4.1). How this three-level hierarchy is applied depends on the situation. For example, when analysing a web forum, the corpus may be a specific thread, in which case the documents would be its posts. Or, with a broader scope of analysis, the corpus may be the entire forum, each thread being a document. Further, the web forum could be the corpus, and each post could be a document. It is therefore important to define explicitly the hierarchy levels.

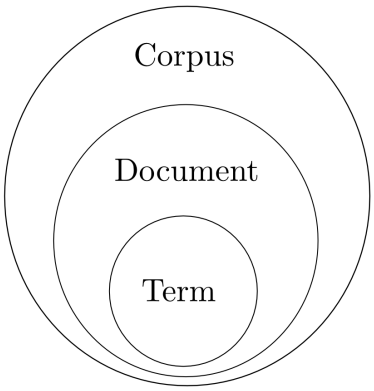


Figure 4.1: Hierarchy in the analysis of texts. A term is a word, a document is a collection of terms, and a corpus is a collection of documents. *Source: Blei et al. [17, p. 995].*

The basic task in IR is to assign a term or a group of terms that can appropriately identify a document in a corpus [40, p. 52]. This process is called *(document) indexing* [40, p. 52]. It can be performed manually by trained experts, but this is not feasible for large amounts of texts [40, p. 52]. In the latter case, an automatic indexing algorithm is performed [40, p. 52]. The process of indexing is aimed at producing the best *index terms*. An index term is "a (document) word whose semantics helps in remembering the document's main themes" [15, p. 24]. The specificity of an index term with respect to a document is measured by its *weight* [15, p. 24].

For the case of a vocabulary $\mathcal{V}$ with t terms, and a corpus $\mathcal{C}$ with N documents, let k_i, $\mathrm{i} \in \{1, \ldots, t\}$ be an index term and d_j, $j \in \{1, \ldots, N\}$ be a document [15, p. 24]. The weight of term k_i in document d_j is denoted as $w_{\mathrm{i},j} \geq 0$, with $w_{\mathrm{i},j} = 0$ if term k_i does not appear in document d_j [15, p. 25]. This allows to assign to each document its *index term vector* $\vec{\mathrm{d}}_j = (w_{1,j}, \ldots, w_{t,j})$ [15, p. 25]. The projection function g_i is defined for each $\mathrm{i} \in \{1, \ldots, t\}$, which returns the weight associated with the term k_i: $g_\mathrm{i}(\vec{\mathrm{d}}_j) = w_{\mathrm{i},j}$ [15, p. 25].

[4]The *bag of words* assumption is used: the order of words is deemed to be irrelevant [17, p. 994]. This is not true from the semantics viewpoint, but this approximation still allows acceptable retrieval of relevant documents.

A formal definition of an IR model is given by Baeza-Yates and Ribeiro-Neto:

> "An information retrieval model is a quadruple $[\mathbf{D}, \mathbf{Q}, \mathcal{F}, R(q_\mathrm{i}, \mathrm{d}_j)]$ where
>
> 1. $\mathbf{D}$ is a set composed of logical views (or representations) for the documents in the collection.
>
> 2. $\mathbf{Q}$ is a set composed of logical views (or representations) for the user information needs. Such representations are called queries.
>
> 3. $\mathcal{F}$ is a framework for modeling document representations, queries, and their relationships.
>
> 4. $R(q_\mathrm{i}, \mathrm{d}_j)$ is a ranking function which associates a real number with a query $q_\mathrm{i} \in \mathbf{Q}$ and a document representation $\mathrm{d}_j \in \mathbf{D}$. Such ranking defines an ordering among the documents with regard to the query q_i." [15, p. 23]

$\mathbf{D}$ and $\mathbf{Q}$ denote the set of available documents and possible queries respectively. The framework manages the representations of documents and queries, and how to assign relevant documents to a given query. Finally, the ranking function orders the relevant document found, in decreasing order of similarity according to the prompted query. There are three main categories of IR models: boolean models, vector models and probabilistic models [15, pp. 20–21]. The frameworks associated to each of these are shown in table 4.1.

	Model framework	
IR model category	Objects representations	Operation on objects
Boolean model	boolean sets	standard operations on sets
Vector model	vectorial space	standard linear algebra operations
Probabilistic model	sample space	standard operations on probabilities

Table 4.1: Description of frameworks for the main categories of IR models. In particular, the vector cosine is important for vector models, as well as Bayes' theorem for probabilistic models. *Source: Modern Information Retrieval [15, pp. 23–24]*

4.1.1 Boolean models

Boolean models assume binary weights, i.e. $w_{\mathrm{i},j} \in \{0, 1\}$: a word k_i is either present ($w_{\mathrm{i},j} = 1$) or absent ($w_{\mathrm{i},j} = 0$) from a document d_j [15, p. 26]. For a query $[q = CAR \wedge (MOTOR \vee \neg DIESEL)]$, the disjunctive normal form can be written as $[\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$, where each element — a *conjunctive component* — is a vector of binary weights associated to the tuple (CAR,$MOTOR$,$DIESEL$). If the vocabulary is $\mathcal{V} = \{CAR,MOTOR,DIESEL,PETROL\}$, then the document d ="$CAR\ MOTOR\ PETROL$" has an index term vector $\vec{\mathrm{d}} = (1,1,0,1)$ with respect to vocabulary $\mathcal{V}$. Furthermore, its representation restricted to the tuple (CAR, $MOTOR$, $DIESEL$) is $\vec{\mathrm{d}}' = (1,1,0)$. Since $(1,1,0) \in \vec{q}_{dnf}$, the document d is retrieved as a consequence of query q. In simpler words, as d contains the word $MOTOR$, the logical expression $MOTOR \vee \neg DIESEL$ is true. Since d

also contains the word CAR, the overall logical expression $CAR \wedge (MOTOR \vee \neg DIESEL)$ is true: d matches the query q.

The similarity of document d_j with query q is defined as [15, p. 26]:

$$sim(\mathrm{d}_j, q) = \begin{cases} 1 & \text{if } \exists \vec{q}_{cc} | (\vec{q}_{cc} \in q_{\mathrm{d}nf}) \wedge (\forall k_\mathrm{i}, g_\mathrm{i}(\vec{\mathrm{d}}_j) = g_\mathrm{i}(\vec{q}_{cc})) \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where $\vec{q}_{cc}$ is conjunctive component of $\vec{q}_{\mathrm{d}nf}$. The document d_j is deemed relevant with respect to query q if $sim(\mathrm{d}_j, q) = 1$, or not relevant if $sim(\mathrm{d}_j, q) = 0$. The boolean model is simple, but tends to retrieve too many documents or too few, because of its binary nature.

4.1.2 Vector models

Vector models assume positive real weights $w_{\mathrm{i},j} \in \mathbb{R}_+$, which allows a more precise ranking of retrieved documents. A *query vector* $\vec{q} = (w_{1,q}, \ldots, w_{t,q})$ is assigned to a query q, and a document d_j is still represented by its index term vector $\vec{\mathrm{d}}_j = (w_{1,j}, \ldots, w_{t,j})$. Both the query q and the documents d_j are therefore represented as t-dimensional vectors $\vec{q} \in \mathbb{R}_+^t$ and $\vec{\mathrm{d}}_j \in \mathbb{R}_+^t$ respectively. The similarity of document d_j with query q is measured by the vector cosine between $\vec{q}$ and $\vec{\mathrm{d}}_j$:

$$sim(\mathrm{d}_j, q) = \cos(\vec{\mathrm{d}}_j, \vec{q}) = \frac{\vec{\mathrm{d}}_j \cdot \vec{q}}{\|\vec{\mathrm{d}}_j\| \, \|\vec{q}\|} = \frac{\sum_{\mathrm{i}=1}^{t} w_{\mathrm{i},j} w_{\mathrm{i},q}}{\sqrt{\sum_{\mathrm{i}=1}^{t} w_{\mathrm{i},j}^2} \sqrt{\sum_{\mathrm{i}=1}^{t} w_{\mathrm{i},q}^2}} \tag{4.2}$$

As the weights are positive, and using the Cauchy-Schwarz inequality, it arises that $0 \leq sim(\mathrm{d}_j, q) \leq 1$. Therefore, documents that only matches *partially* the query may be retrieved. The documents are ordered according to their similarity, and a threshold may be used to determine the relevant ones, above a required level of similarity.

The *tf-idf* scheme is a widely used representation of documents, which balances the *term frequency* with the *inverse document frequency*. The frequency of term k_i in document d_j is denoted by $freq_{\mathrm{i},j}$, and the normalized term frequency is defined as:

$$\mathrm{tf}_{\mathrm{i},j} = \frac{freq_{\mathrm{i},j}}{\max_l freq_{l,j}} \tag{4.3}$$

thus $0 \leq \mathrm{tf}_{\mathrm{i},j} \leq 1$: if term k_i does not appear in document d_j, then $\mathrm{tf}_{\mathrm{i},j} = 0$. If it is its most frequent term, then $\mathrm{tf}_{\mathrm{i},j} = 1$. The more frequent is term k_i in a document, the higher is its term frequency. If N is the total number of documents ($N = |\mathcal{C}|$) and n_i is the number of documents in which term k_i appear, the inverse document frequency is written:

$$\mathrm{idf}_\mathrm{i} = \log \frac{N}{n_\mathrm{i}} \tag{4.4}$$

thus $0 \leq \text{idf}_\text{i} \leq \log N$. If term k_i appears in all the corpus' documents, then $\text{idf}_\text{i} = \log \frac{N}{N} = 0$. On the contrary, if term k_i appears in only one document, then $\text{idf}_\text{i} = \log N$. Therefore, the less documents term k_i belongs to, the higher its inverse document frequency is. The overall weight is computed as the product of the normalized term frequency and the inverse document frequency:

$$w_{\text{i},j} = \text{tf-idf}_{\text{i},j} = \text{tf}_{\text{i},j} \times \text{idf}_\text{i} = \frac{freq_{\text{i},j}}{\max_l freq_{l,j}} \times \log \frac{N}{n_\text{i}} \tag{4.5}$$

The index term vector for the query has weights which are computed according to:

$$w_{\text{i},q} = \text{tf-idf}_{\text{i},q} = \left(0.5 + \frac{0.5 freq_{\text{i},q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_\text{i}} \tag{4.6}$$

The tf-idf scheme considers the term frequency in order to rank the relevant documents, and the role of inverse document frequency s to distinguish relevant from non relevant documents.

4.1.3 Probabilistic models and Bayesian networks

The basic probabilistic model is based on the following assumption:

> "*Assumption (Probabilistic Principle)* Given a user query q and a document d_j in the collection, the probabilistic model tries to estimate the probability that the user will find the document d_j interesting (i.e., relevant). The model assumes that this probability of relevance depends on the query and the document representations only. Further, the model assumes that there is a subset of all documents which the user prefers as the answer set for the query q. Such an *ideal* answer set is labeled R and should maximize the overall probability of relevance to the user. Documents in the set R are predicted to be *relevant* to the query. Documents not in this set are predicted to be *non-relevant*." [15, p. 31].

However, the basic probabilistic model reviewed by Baeza-Yates and Ribeiro-Neto is different from the model used in this work, the LDA model, and therefore is not considered here. In return, Bayesian networks are introduced as a central modeling tool.

A Bayesian network is represented as a directed acyclic graph (DAG) that represents a network of causal relationships between random variables [**?**, p. 49]. The set of parent nodes a of a given node are assumed to be its direct causes, and the nodes without parents are called *roots* [**?**, p. 49]. Figure 4.2 shows an example of such a network.

One of the advantages of Bayesian networks consists in simplified joint probability distribution. Consider the joint probability distribution $Pr(x_1, x_2, x_3, x_4, x_5)$ in the context of

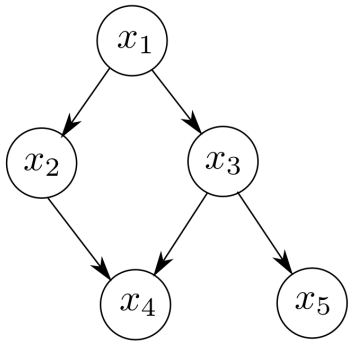


Figure 4.2: Example of a Bayesian network. *Source: reproduced from* Modern Information Retrieval *[15, p. 49].*

figure 4.2. It can be written as:

$$\begin{aligned} Pr(x_1, x_2, x_3, x_4, x_5) &= Pr(x_1)\frac{Pr(x_1, x_2, x_3)}{Pr(x_1)}\frac{Pr(x_1, x_2, x_3, x_4, x_5)}{Pr(x_1, x_2, x_3)} \\ &= Pr(x_1)Pr(x_2, x_3|x_1)Pr(x_4, x_5|x_1, x_2, x_3) \end{aligned} \tag{4.7}$$

The network structure is now used, through the causal Markov condition, which states that any node in a Bayesian network is conditionally independent[5] of its nondescendants, given its parents. The nodes x_2 and x_3 are both children of the same parent node x_1: neither x_2 is a descendant of node x_3, nor vice versa. Therefore, x_2 and x_3 are conditionally independent given x_1, so:

$$Pr(x_2, x_3|x_1) = Pr(x_2|x_1)Pr(x_3|x_1) \tag{4.8}$$

Analogously:

$$Pr(x_4, x_5|x_1, x_2, x_3) = Pr(x_4|x_2, x_3)Pr(x_5|x_3) \tag{4.9}$$

By 4.8 and 4.9, equation 4.7 becomes:

$$Pr(x_1, x_2, x_3, x_4, x_5) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_1)Pr(x_4|x_2, x_3)Pr(x_5|x_3) \tag{4.10}$$

Therefore, the joint distribution is equal to the products of each node probabilities conditioned by their parents' value.

4.2 The latent Dirichlet allocation

The *latent Dirichlet allocation*[6] (LDA) is a generative probabilistic model proposed by Blei et al. in 2003 [17] that represents text documents as vectors of topic probabilities, instead of

[5]Two random variables X and Y are said to be conditionally independent given a random variable Z if $Pr(X, Y|Z) = Pr(X|Z)Pr(Y|Z)$, which is equivalent to $Pr(X|Y, Z) = Pr(X|Z)$.

[6]This section is based on the paper by Blei et al., *Latent Dirichlet Allocation* [17].

vectors of term weights (as in the tf-idf scheme). Since the number of topics is usually much smaller than the number of index terms, it achieves a significant dimensionality reduction with respect to traditional vector models. Furthermore, it outperforms previous probabilistic models (such as the unigram, mixture of unigrams and pLSI models; [17]). Finally, it is a modular model that can be modified and adapted to specific needs, and which allows the representation of documents not included in the initial training set. These advantages justify the election of LDA in the context of this work.

Below, the core generative process of LDA is first defined. Then, the most important distributions are deduced, and the inference of topic probabilities , as well as the estimation of the model's parameters, are discussed. Finally, the use of LDA within this work's context is addressed.

4.2.1 Generative process

The following notation is used for terms, documents and the corpus [17, p. 995]:

- for a vocabulary $\mathcal{V}$ indexed by $\{1, \ldots, V\}$, a term is modeled as a V-dimensional vector w such that $w^v = 1$ (v is the term's index in $\mathcal{V}$) and $w^u = 0 \, \forall u \neq v$. Therefore, w is a vector filled with zeros, except in its v^{th} component. The Kronecker delta notation is used: $w^u = \delta_{u,v}$.
- a document is modeled as a sequence of N terms $\mathbf{w} = (w_1, \ldots, w_N)$.
- a corpus is modeled as a collection of M documents $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$.

The terms, documents and corpus are the basic units of analysis used throughout the model.

Given a predefined number of topics k, the latent Dirichlet allocation is a a three-level Bayesian model with parameters α and β. The former is a k-dimensional vector $\alpha \in \mathbb{R}_+^k$, while the latter is a $k \times V$ matrix $\beta \in \mathcal{M}_{k \times V}(\mathbb{R}_+)$, such that $\sum_{j=1}^{V} \beta_{\mathrm{i},j} = 1 \, \forall \mathrm{i} \in \{1, \ldots, k\}$. Blei et al. define the core generative process as follows:

1. "Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words words w_n:
   a Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   b Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n." [17, p. 996]

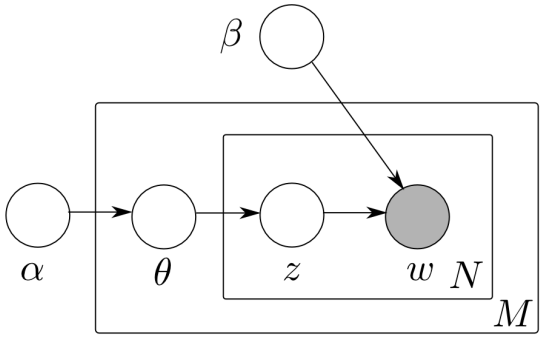


Figure 4.3: Plate notation of the LDA model. For each document, the topics probabilities θ are sampled. Then, for each term w_n, its topic z_n is first determined, and then the term is sampled. *Source: reproduced from Blei et al.,* Latent Dirichlet Allocation *[17].*

The generative process states that the document size is first sampled (once per document). As this sampling is independent of the rest, no attention is paid to it in the sequel. Then, the vector of topic probabilities θ is sampled (once per document), which in turn determines the set of topic assignments $\mathbf{z} = (z_1, \ldots, z_N)$ (N samplings per document). Finally, the document $\mathbf{w} = (w_1, \ldots, w_N)$ is generated from $\mathbf{z}$ and β (N samplings per document). The overall process is represented in figure 4.3.

The vector of topics probabilities θ is a k-dimensional vector in the $(k-1)$-simplex $\mathcal{S}_{k-1} \subset \mathbb{R}^k_+$. As $\theta \in \mathcal{S}_{k-1}$, then $\sum_{\mathrm{i}=1}^{k} \theta_{\mathrm{i}} = 1$, hence each component θ_{i} may be interpreted as a probability. θ is sampled from a Dirichlet distribution on $\mathcal{S}_{k-1}$ with parameter $\alpha \in \mathbb{R}^k_+$, which has the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{\mathrm{i}=1}^{k} \alpha_{\mathrm{i}})}{\prod_{\mathrm{i}=1}^{k} \Gamma(\alpha_{\mathrm{i}})} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \quad \text{[17, p. 996]} \tag{4.11}$$

The topic assignment z_n is modeled as a k-dimensional vector filled with zeros, except for one component i where $z_n^{\mathrm{i}} = 1$: in this case, the i^{th} topic has been assigned, which occurs with probability θ_{i}. z_n is therefore sampled from a multinomial distribution with parameter θ. A term w_n is a V-dimensional vector filled with zeros, except for one component j where $w_n^j = 1$: given that the i^{th} topic has been assigned to w_n, this occurs with probability $\beta_{\mathrm{i},j} = Pr(w^j = 1 | z^{\mathrm{i}} = 1)$. w_n is thus sampled from a multinomial distribution with parameter $\beta_{\mathrm{i},\bullet}$.

4.2.2 Joint and marginal distributions

The joint distribution of a topic mixture θ, a set of topic assignments $\mathbf{z}$ and a document $\mathbf{w}$ is:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad \text{[17, p. 996]} \tag{4.12}$$

This joint distribution is a consequence of the LDA Bayesian network topology (figure 4.4), with its consequent set of conditional independencies (discussed previously in section 4.1.3). The marginal distribution of a document $\mathbf{w}$ is obtained by integrating over θ and summing over z [17, p. 997]:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\Big(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\Big)\mathrm{d}\theta \tag{4.13}$$

As discussed above, $p(z_n|\theta) = \theta_{\mathrm{i}}$ where $z_n^{\mathrm{i}} = 1$, and $p(w_n|z_n,\beta) = \beta_{\mathrm{i},s}$ where $w_n^s = 1$. Using equation 4.11, the marginal distribution of a document takes the explicit form:

$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_{\mathrm{i}=1}^{k}\alpha_{\mathrm{i}})}{\prod_{\mathrm{i}=1}^{k}\Gamma(\alpha_{\mathrm{i}})}\int\Big(\prod_{\mathrm{i}=1}^{k}\theta_{\mathrm{i}}^{\alpha_{\mathrm{i}}-1}\Big)\Big(\prod_{n=1}^{N}\sum_{\mathrm{i}=1}^{k}\prod_{j=1}^{V}(\theta_{\mathrm{i}}\beta_{\mathrm{i},j})^{w_n^j}\Big)\mathrm{d}\theta \tag{4.14}$$

Finally, the probability of a corpus $\mathcal{D}$ is obtained as the product of the marginal probabilities of single documents [17, p. 997]:

$$p(\mathcal{D}|\alpha,\beta) = \prod_{\mathrm{d}=1}^{M}\int p(\theta_{\mathrm{d}}|\alpha)\Big(\prod_{n=1}^{N_{\mathrm{d}}}\sum_{z_{\mathrm{d}n}} p(z_{\mathrm{d}n}|\theta_{\mathrm{d}})p(w_{\mathrm{d}n}|z_{\mathrm{d}n},\beta)\Big)\mathrm{d}\theta_{\mathrm{d}} \quad \text{[17, p. 997]} \tag{4.15}$$

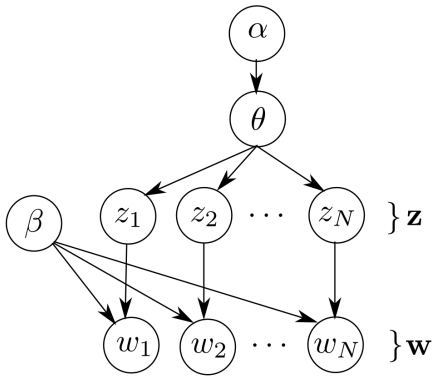


Figure 4.4: "Uncompressed" view of the Bayesian network associated to a single document $\mathbf{w}$. First, the topics probabilities θ are sampled. Then, for each term w_n, its topic z_n is sampled from a multinomial distribution with parameter θ, and w_n is sampled from a multinomial distribution with parameter $\beta_{\mathrm{i},\bullet}$, with $z_n^j = \delta_{\mathrm{i}j}$. *Source: Blei et al.,* Latent Dirichlet Allocation *[17].*

4.2.3 Inference and parameters estimation

The inference of the topic probabilities assigned to each document $\mathbf{w}$, and the estimation of the parameters α and β of the model, are realized in two successive steps. Altogether, they constitute a *variational EM procedure* [17, p. 1005]. In the first step (called the *E-step*), the generative process of the topic probabilities θ and the topic assignments $\mathbf{z}$ is replaced by a

simpler one. This allows, for each document, the computation of a lower bound of the actual log likelihood of the document, which is indeed intractable. In the second step (called the *M-step*), a lower bound of the log likelihood of the whole corpus is maximized with respect to α and β, thus estimating the parameters of the model. Both steps are discussed below.

E-step: variational inference [7]

The inferential problem consists in the calculation of the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{4.16}$$

However, the normalization term $p(\mathbf{w}|\alpha, \beta)$ is intractable to compute (equation 4.14) due to the coupling between θ and β. Therefore, the posterior is approximated by the variational distribution:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{4.17}$$

where a simplified generative process for θ and $\mathbf{z}$ is assumed (figure 4.5).

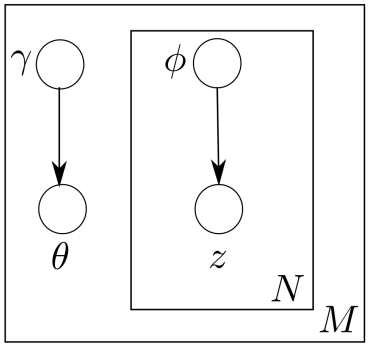


Figure 4.5: Simplified generative process for θ and $\mathbf{z}$. θ is sampled from a Dirichlet with parameter γ, and z from a multinomial with parameter ϕ. *Source: reproduced from Blei et al.,* Latent Dirichlet Allocation *[17].*

The logarithm of the marginal distribution of a document $\mathbf{w}$ is written:

$$\begin{aligned}
\log p(\mathbf{w}|\alpha, \beta) &= \log \int \sum_{z} p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) \mathrm{d}\theta \\
&= \log \int \sum_{z} \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) q(\theta, \mathbf{z}|\gamma, \phi)}{q(\theta, \mathbf{z}|\gamma, \phi)} \mathrm{d}\theta \\
&= \log E_q \left[\frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z})} \right]
\end{aligned} \tag{4.18}$$

[7] [17], pp. 1003–1005 and pp. 1019–1021.

Here, the Jensen inequality is used, which states that for a random variable X and a convex function f, the inequality $f(E[X]) \leq E[f(X)]$ holds. As log is a concave function, then the inequality is reversed so $\log(E[X]) \geq E[\log(X)]$. Therefore, equation 4.18 becomes:

$$\log p(\mathbf{w}|\alpha,\beta) \geq E_q\left[\log \frac{p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)}{q(\theta,\mathbf{z})}\right] \tag{4.19}$$

The right hand side is denoted $\mathcal{L}(\gamma,\phi;\alpha,\beta)$ and provides a lower bound for the log likelihood of a document $\mathbf{w}$. It can be shown that the maximization of $\mathcal{L}(\gamma,\phi;\alpha,\beta)$ yields the following algorithm:

initialize $\phi^0_{n\mathrm{i}} := 1/k \ \forall \mathrm{i}, n$
initialize $\gamma_\mathrm{i} := \alpha_\mathrm{i} + N/k \quad \forall \mathrm{i}$
repeat
  for $n = 1$ **to** N
    for $\mathrm{i} = 1$ **to** k
      $\phi^{t+1}_{n\mathrm{i}} := \beta_{\mathrm{i}w_n} \exp(\Psi(\gamma^t_\mathrm{i}) - \Psi(\sum_{j=1}^k \gamma_j))$
    normalize ϕ^{t+1}_n to sum to 1.
  $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi^{t+1}_n$
until convergence of $\mathcal{L}(\gamma,\phi;\alpha,\beta)$

where $\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$, which is computable with Taylor approximations. When the algorithm stops, document-specific optimal values γ^* and ϕ^* are obtained, which are indeed functions of $\mathbf{w}$. In particular, Dirichlet parameters $\gamma^*(\mathbf{w})$ provide a representation of a document in the topic simplex: $\gamma^*(\mathbf{w})$ *parameters are the representation of document* $\mathbf{w}$ *in the LDA model*. Also, the variational distribution $q(\theta,\mathbf{z}|\gamma^*(\mathbf{w}),\phi^*(\mathbf{w}))$ can be viewed as an approximation of the posterior distribution $p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$.

M-step: parameter estimation [8]

The log likelihood of the whole corpus $\mathcal{C} = \{\mathbf{w}_1,\ldots,\mathbf{w}_M\}$ is equal to:

$$\ell(\alpha,\beta) = \sum_{\mathrm{d}=1}^{M} \log p(\mathbf{w}_\mathrm{d}|\alpha,\beta) \tag{4.20}$$

By equation 4.19, a lower bound for $\ell(\alpha,\beta)$ is:

$$\ell(\alpha,\beta) \geq \sum_{\mathrm{d}=1}^{M} E_q\left[\log \frac{p(\theta_\mathrm{d},\mathbf{z}_\mathrm{d},\mathbf{w}_\mathrm{d}|\alpha,\beta)}{q(\theta_\mathrm{d},\mathbf{z}_\mathrm{d})}\right] \tag{4.21}$$

The maximization of the lower bound yields the following multinomial parameters:

$$\beta_{\mathrm{i}j} \propto \sum_{\mathrm{d}=1}^{M}\sum_{n=1}^{N_\mathrm{d}} \phi_{\mathrm{d}n\mathrm{i}} w^j_{\mathrm{d}n} \tag{4.22}$$

[8] [17], pp. 1005–1006 and pp. 1021–1022.

where the scaling arises from $\sum_{j=1}^{V} \beta_{ij} = 1 \quad \forall i = \{1, \ldots, k\}$. The Dirichlet parameters come from the zeros of:

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_i} = M(\Psi(\sum_{j=1}^{k} \alpha_j) - \Psi(\alpha_i)) + \sum_{d=1}^{M} \big(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})\big) \tag{4.23}$$

with:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{M} \left(\log \Gamma(\sum_{j=1}^{k} \alpha_j) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} \big((\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))\big) \right) \tag{4.24}$$

4.2.4 Use of LDA in this work

A useful concept is that of *topic profile*, which is associated to the LDA vector representation of a text document. Let $\mathbf{p}$ be a LDA vector and k the number of topics considered. Therefore, $\mathbf{p}$ is a vector of k non negative components, which satisfy the condition $\sum_{i=1}^{k} (\mathbf{p})_i = 1$. The i^{th} component $(\mathbf{p})_i$ is called the *topic weight* of topic i, and represents the relative strength of the topic within the document. The topic profile associated to $\mathbf{p}$ is the sequence of topics weights $\big((\mathbf{p})_1, \ldots, (\mathbf{p})_k\big)$, and its graphical representation is obtained from drawing the sequence versus the topic indexes $(1, \ldots, k)$ (figure 4.6).

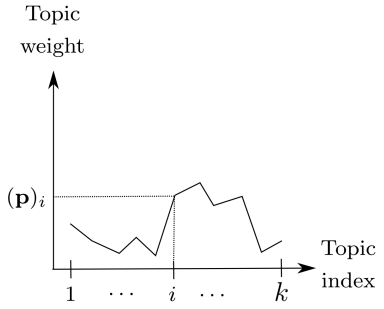


Figure 4.6: Graphical representation of the topic profile of a LDA vector $\mathbf{p}$.

Chapter 5

Description of the models

Insofar, theoretical aspects of general diffusion, decision making and information retrieval have been discussed. Diffusion is conceived as a collective process that operates over a network structure. Two models of choice, the LCA model and the logit model, have been analysed. Both are probabilistic in nature, the first through a set of stochastic differential equations, the second by the means of explicitly computed probabilities. Information retrieval is the process of representing and finding text documents among large collections. This theoretical framework is the basis for a set of implemented models in the context of this work, which seek to simulate the dynamics of a real web forum. It provides the information modeling, the decision-making as the underlying diffusion mechanism, and the overall diffusion perspective.

The implemented models operate within a general model of the forum users activity. This framework specifies a forum structure, as well as users' interactions with it. Among user actions, navigation actions are distinguished from content actions. Navigation actions define users' interactions with the forum structure, while content actions set the interactions with its contents. In this chapter, the general model is presented, and then four instances are discussed, with their respective decision-making mechanisms.

5.1 Forum-Agent system framework

As noted in section 2.4, the structure of the network over which a diffusion process takes place is relevant. Within this work's scope, the underlying structure is a simplified version of the web forum's threads tree. The diffusion process does not occur over a graph of users. Rather, the users' interaction graph is a consequence of the diffusion process, as discussed in the next chapter (Chapter 6). In this section, the simplified forum representation is described, where agents navigate and interact within this virtual media.

5.1.1 Forum structure

The forum is modeled as a three level structure. The highest level is the *forum level*, then comes the *thread level*, and finally the lowest level is the *post level*. The forum is modeled as a list of threads, and each thread in turn as a list of posts. Let N be the number of current threads in the forum, so the threads are denoted as $T_1, \ldots, T_N$. The current posts of thread T_i are modeled as LDA topic weights vectors (Chapter 4, section 4.2.4), which are denoted $\mathbf{p}_{\mathrm{i},1}, \ldots, \mathbf{p}_{\mathrm{i},n_\mathrm{i}}$. The publication time of the j^{th} post in thread T_i is $t(\mathbf{p}_{\mathrm{i},j})$. Furthermore, the posts of each thread are ordered by increasing publication time, so the following ordering holds:

$$t(\mathbf{p}_{\mathrm{i},1}) \leq \ldots \leq t(\mathbf{p}_{\mathrm{i},j-1}) \leq t(\mathbf{p}_{\mathrm{i},j}) \leq t(\mathbf{p}_{\mathrm{i},j+1}) \leq \ldots \leq t(\mathbf{p}_{\mathrm{i},n_\mathrm{i}}) \, \forall \mathrm{i} \in \{1, \ldots, N\} \tag{5.1}$$

On the other hand, the publication time of the i^{th} thread $t(T_\mathrm{i})$ is defined as the publication time of its last post added, so $t(T_\mathrm{i}) = t(\mathbf{p}_{\mathrm{i},n_\mathrm{i}})$. In contrast with the ordering of a thread's posts, the threads of the forum are ordered by decreasing publication time:

$$t(T_1) \geq \ldots \geq t(T_{\mathrm{i}-1}) \geq t(T_\mathrm{i}) \geq t(T_{\mathrm{i}+1}) \geq \ldots \geq t(T_N) \tag{5.2}$$

The reason for the previous orderings is rooted in the actual behavior of the forum when browsed. The more recent threads appear first, while the oldest posts appear at the top within each thread. Nonetheless, this structure (summarized in figure 5.1) is a simplification of the real forum structure. In effect, threads do not constitute a single list, but are grouped into discussion topics (not to be mistaken with LDA topics). In the Plexilandia forum, there are actually six of these topics: amplifiers, effects, synthesizers, lutherie, professional audio, general discussion [11]. Moreover, when one topic is selected, all of its threads do not appear at once, but by a fixed amount at a time. In the same way, a thread's posts do not appear at once either. It is important to keep in mind those simplifications for the later analysis of results.

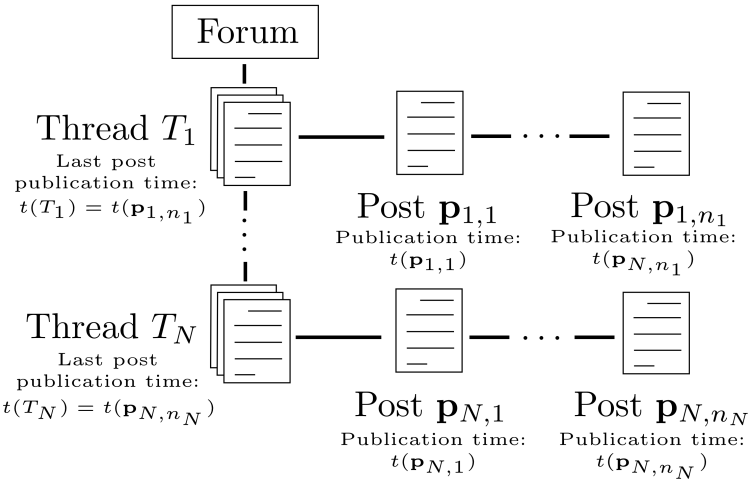


Figure 5.1: Representation of the modeled forum structure, which is actually a simplification of the real structure.

5.1.2 Users behavior

Users perform actions while browsing the forum, such has choosing threads, reading posts and publishing new contents. These are called *user actions*. Each of these actions has a formal rule that defines it, giving rise to *action rules*. Below, both *user actions* and *action rules* are discussed. An integrated theory of mind can be found in [13], and experiments of information foraging users in online settings are described in [33, 34].

User actions

Users are fundamentally defined by two state vectors : a *preference* vector, and a *social image* vector. The preference vector represents the user's preferred topics of conversation, while the social image vector represents his fellows' perceptions about him, again in terms of topics. Both vectors are therefore modeled as LDA vectors: the preference vector is denoted by μ, and the social image vector is denoted by ν. In order to generate dynamics, the user performs *content actions*. There are two such actions: `read-a-post`, and `publish-a-post`. When a user reads a post, it is kept in memory. A user's memory therefore contains the last post that has been read by the user. Moreover, the reading affects his preference, and the preference vector μ changes. On the other hand, when a user publishes a post, a new post is generated and his social image evolves too: the social image ν now changes. A basic sequence of reading-posting is illustrated in figure 5.2, for a clearer view of the process.

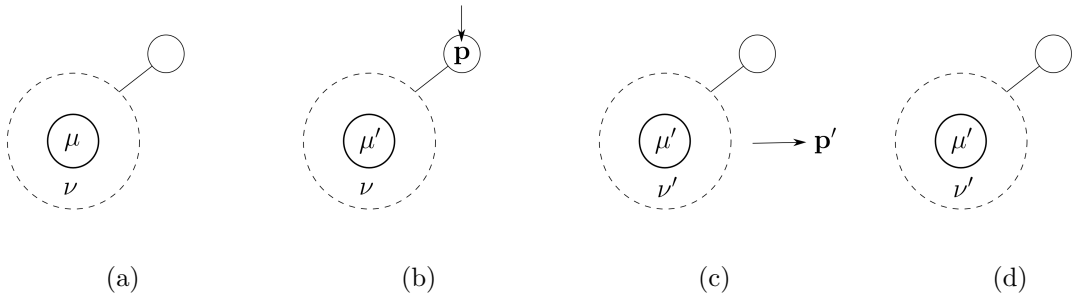


Figure 5.2: Example of a reading and posting sequence. The user has a preference vector μ, and a social image vector ν. Subfigure (a): initial state of the user. Subfigure (b): the user reads a post $\mathbf{p}$ (which is kept in memory), so his preference is modified. Subfigure (c): the user replies to $\mathbf{p}$ by publishing a new post $\mathbf{p'}$. His social image is modified now. Subfigure (d): final state of the user.

There are also four *navigation actions*, which govern the movements of the users within the forum's structure. The first one is `begin-session`, when the user accesses to the forum web site. The second one is `choose-a-thread`, which occurs if the user selects a thread of discussion, among all available threads. The third navigation action is `choose-a-post`, where the user selects a post, among all published posts in a given thread. Finally, `end-session` happens when the user exits from the forum web site.

A *session* is a sequence of interaction and navigation actions, starting with a `begin-session` action, ending with a `end-session` action, and with no start or end of session in between. One such a sequence may be, for example:

```
begin-session
choose-a-thread
choose-a-post
read-post
choose-a-thread
choose-a-post
read-post
publish-a-post
end-session
```

However, not all sequences are possible. A `begin-session,publish-a-post,end-session` sequence is clearly not admissible, as the user publishes a post without even choosing a thread before. Thus, an action diagram (figure 5.3) defines the set of possible transitions between actions. The diagram contains five nodes: (ON) for `begin-session`, (T) for `choose-a-thread`, (R) for the combination of `choose-a-post` and `read-a-post`, (P) for `publish-a-post` and (OFF) for `end-session`. *Each allowed transition has an associated probability of occurrence, as well as an exponentially distributed time of duration.* Also, it must be noted that a user reads or publishes a post in the last thread he has chosen.

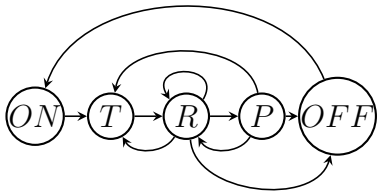


Figure 5.3: User actions diagram.

Action rules

Six *user actions* were defined above and classified as *navigation actions* (`begin-session`, `choose-a-thread`, `choose-a-post`, `end-session`) or *content actions* (`read-a-post`, `publish-a-post`). Though, the effects of these actions still must be defined. Therefore, an *action rule* is assigned to each action, and more precisely, *navigation rules* and *content rules* are associated with navigation actions and content actions respectively.

Starting with the latter kind of action, three interaction rules are set, inspired in DeGroot theory of belief updating [12]:

$$\mu' = c_\mu \mathbf{p} + (1 - c_\mu)\mu \quad \textit{read-a-post} \text{ content rule} \tag{5.3}$$

$$\mathbf{p}' = c_\pi \mathbf{p} + (1 - c_\pi)\mu' \quad \textit{publish-a-post} \text{ content rule 1} \tag{5.4}$$

$$\nu' = c_\nu \mathbf{p}' + (1 - c_\nu)\nu \quad \textit{publish-a-post} \text{ content rule 2} \tag{5.5}$$

so two rules are assigned to `publish-a-post`. Each rule has an associated parameter (namely, c_μ, c_ν and c_π) whose range is a subset of the $[0, 1]$ unit interval, and for a specific model instance, all users have the same parameter values. Also, distinct values of those may bring very different dynamics. For example, if $c_\mu = 0$ and $c_\pi = 0$, the users maintain a fixed preference, and post exclusively according to it. On the contrary, if $c_\mu = 1$ and $c_\pi = 1$, the users change of preference each time they read a post, and they post the last content they read.

With respect to navigation actions, no action rules are assigned to `begin-session` and `end-session`. Indeed, these are utility actions which are mainly symbolic, in the sense that they help to delimit different sessions, but with no real effect. On the other hand, the rules for `choose-a-thread` and `choose-a-post` are particular to the specific implementation of the general model. A summary of user actions is found in table 5.1.

User action	Action type	Diagram node	Action rule
`read-a-post`	content	R	eq. 5.3
`publish-a-post`	content	P	eqs. 5.4 and 5.5
`begin-session`	navigation	ON	none
`choose-a-thread`	navigation	T	model-specific
`choose-a-post`	navigation	R	model-specific
`end-session`	navigation	OFF	none

Table 5.1: Summary of the six user actions. Two of them are content actions, four are navigation actions.

5.2 Decision-making models of OSN users

To this moment, the general model has been described through its two main components, *i.e.* the forum structure and the user actions. The latter were further classified into two categories: navigation actions, and content actions. A total of six user actions were mentioned, but only four were defined. The two remaining user actions — `choose-a-thread` and `choose-a-post` — are therefore specific to model implementations, which are now discussed below.

5.2.1 Main model: LCA-based and random

The LCA-based model is the main model of this work, as one of the hypotheses to test is to verify whether it produces a better quality in contents and graph generation (Chapter 1, section 1.2). It is based on the leaky competing accumulator model by Usher and McClelland [44], which is a perceptual choice model: its goal is to model how decisions are taken from sensory neuronal inputs. In the case of the model, the inputs are not sensory but rather more

abstract. Indeed, the processing of text contents occur in a higher level than the processing of a visual stimulus. Henceforth, in order to provide a reasonable vector of inputs to the LCA model, the logit model is invoked.

The *choose-a-thread* user action

Suppose that at instant t, a user u, with preference μ, has to choose between N threads $T_1, \ldots, T_N$ in the forum. The i^{th} thread has a publication time $t(T_\mathrm{i})$ and a first post $\mathbf{p}_{\mathrm{i},1}$. The deterministic component of the utility obtained by choosing thread T_i is:

$$v_\mathrm{i} = \alpha_\mu \cos(\mu, \mathbf{p}_{\mathrm{i},1}) + \alpha_t \,\mathrm{norm}(t - t(T_\mathrm{i})) \tag{5.6}$$

where cos denotes the vector cosine (Chapter 4, section 4.1.2, equation 4.2), and $\mathrm{norm}(.) : \mathbb{R} \to [0, 1]$ is a normalization function. The vector of logit probabilities $p^{LOGIT} \in \mathbb{R}^N$ then satisfies $\sum_{\mathrm{i}=1}^{N} p_\mathrm{i}^{LOGIT} = 1$, and the probability of choosing the thread T_i is equal to (section 3.2, equation 3.37):

$$p_\mathrm{i}^{LOGIT} = \frac{\mathrm{e}^{v_\mathrm{i}}}{\sum_{j=1}^{N} \mathrm{e}^{v_j}} \tag{5.7}$$

Recalling the LCA model, the inputs vector is defined as $I = \beta p^{LOGIT}$, with $\beta \in \mathbb{R}$ a scalar. Then, it is known (Chapter 3, section 3.1.2, equation 3.33) that the attractor plane intersects with any axis at:

$$y^* = \frac{\sum_{\mathrm{i}=1}^{N} I_\mathrm{i}}{\kappa + (N-1)w} \tag{5.8}$$

If Z is the decision threshold, it is suitable that the attractor plane intersects with the axis near Z. Therefore, the condition $y^* = Z$ is imposed, and as $\sum_{\mathrm{i}=1}^{N} I_\mathrm{i} = \beta$, this leads to:

$$\beta = (\kappa + (N-1)w)Z \tag{5.9}$$

In conclusion, for the LCA-based main model, the choice of a thread is a random process, where a LCA model instance is run over an inputs vector $I = (\kappa + (N-1)w)Z p^{LOGIT}$.

The *choose-a-post* user action

This case is analogous to the choice of a thread, but the utility of choosing the j^{th} post (it is assumed that the user u is browsing thread T_i) incorporates a vector cosine between u's social image ν and the post's author social image $\nu(\mathbf{p}_{\mathrm{i},j})$:

$$v_j = \alpha_\mu \cos(\mu, \mathbf{p}_{\mathrm{i},j}) + \alpha_t \,\mathrm{norm}(t - t(\mathbf{p}_{\mathrm{i},j})) + \alpha_\nu \cos(\nu, \nu(\mathbf{p}_{\mathrm{i},j})) \tag{5.10}$$

so the Logit probabilities are now:

$$p_j^{LOGIT} = \frac{\mathrm{e}^{v_j}}{\sum_{k=1}^{n_\mathrm{i}} \mathrm{e}^{v_k}} \tag{5.11}$$

The remaining analysis is identical than before.

5.2.2 Benchmark: a logit-based random model

The logit is a classical model of discrete choice ([26]; section 3.2), and therefore is a good benchmark for the main LCA-based model. As discussed previously, the main model is run over a logit vector of probabilities. Therefore, the pure logit model is already defined implicitly in the description of the main model. The testing of a pure logit model allows to measure the effect of the inclusion of the LCA model as an upper layer of decision.

The *choose-a-thread* user action

The choice of the thread is ruled by the probabilities given by equation 5.7.

The *choose-a-post* user action

The choice of the post is ruled by the probabilities given by equation 5.11.

5.2.3 Benchmark: a purely random model

Both the main model and the logit-based model make use of available information concerning threads, posts and OSN users. Thus, these models should make better predictions than a purely random model, in which the users choose threads as well as posts completely at random. In order to determine the gain of predictive power of the main model versus a model where no information is exploited, a purely random model is implemented.

The *choose-a-thread* user action

The user chooses the thread T_{i} with probability:

$$p_{\mathrm{i}}^{Rand} = \frac{1}{N} \tag{5.12}$$

The *choose-a-post* user action

The user chooses the post $\mathbf{p}_{\mathrm{i},j}$ with probability:

$$p_{\mathrm{i}}^{Rand} = \frac{1}{n_{\mathrm{i}}} \tag{5.13}$$

5.2.4 Benchmark: a deterministic model

The models described above include, some way or another, a random choice of threads and posts. A deterministic model is therefore incorporated into the framework of analysis, based on the logit vector of probabilities. The model is such that the maximum component of the logit probabilities vector is chosen.

The *choose-a-thread* user action

By equation 5.7, the probability of choosing the thread T_{i} is:

$$p_{\mathrm{i}}^{LOGIT} = \frac{\mathrm{e}^{v_{\mathrm{i}}}}{\sum_{j=1}^{N} \mathrm{e}^{v_j}}$$

Then, the thread that is actually chosen is the thread T_{i^*}, where:

$$\mathrm{i}^* = \mathrm{argmax}\{p_{\mathrm{i}}^{LOGIT}\} \tag{5.14}$$

The *choose-a-post* user action

By equation 5.11, the probability of choosing the post $\mathbf{p}_{\mathrm{i},j}$ is:

$$p_j^{LOGIT} = \frac{\mathrm{e}^{v_j}}{\sum_{k=1}^{n_{\mathrm{i}}} \mathrm{e}^{v_k}}$$

The actually chosen post is then $\mathbf{p}_{\mathrm{i},j^*}$, with:

$$j^* = \mathrm{argmax}\{p_j^{LOGIT}\} \tag{5.15}$$

The models presented above are summarized in table 5.2.

Model code name	Probabilities input vector	Decision-making type	Decision-making mechanism
FreqLCA	Logit	Random	LCA
FreqLogit	Logit	Random	Uniform distribution sampling
FreqRandom	Equiprobable	Random	Uniform distribution sampling
FreqMax	Logit	Deterministic	argmax

Table 5.2: The main LCA-based model (codename: *FreqLCA*) is compared with three benchmark models.

5.3 Discussion

The main LCA-based model of contents choice is compared with three benchmark models: the FreqMax, FreqLogit and FreqRandom models. The FreqRandom is independent from the other models, which are all related to the logit model. The FreqLogit model is a pure logit

model which computes logit probabilities of choice. The FreqMax model is a deterministic model that always chooses the option with greatest probability level. The FreqLCA main model adds a further layer of decision, based on the LCA model, over the logit probabilities. The intuition behind this setting is that the LCA model reproduces better the statistical characteristics of human choice. However, it must be noted that the LCA model has been created for perceptual inputs, while contents decisions are more abstract. There is therefore a possible undesirable circularity in the main model. Indeed, the LCA model — a low-level decision model — is run after the logit model — a high-level decision model.

The models are heavily dependent upon a set of parameters. In particular, c_π, c_μ, c_ν, α_μ, α_ν and α_t are of the uttermost importance. The c parameters influence the posting behavior, while the α parameters are related to choice valuations. Different values of these parameters yield very different dynamics. It is therefore very important to understand the implications of equations 5.3, 5.4, 5.5, 5.6 and 5.10.

Chapter 6

Methodological framework

The models described in chapter 5 have been implemented in the JAVA programming language. The classes belong to package `com.snagroup.diffusion`, a complete JAVA library developed in the context of this work. The package also includes classes that preprocess the data needed by the simulation and analyse the results. The experimental data set consists of a database with posts information, and server web logs. The data availability of posts and logs span different time windows. The models are calibrated during the time period for which both posts and server logs are available, thus between 2009/10/31 and 2010/03/26 (figure 6.1). The simulation time window begins at 2010/03/26 and ends at 2010/08/31, spanning the whole posts availability period after the calibration. As a result of the simulated web forum activity, posts and user interactions graphs are generated. These are compared with available real data.

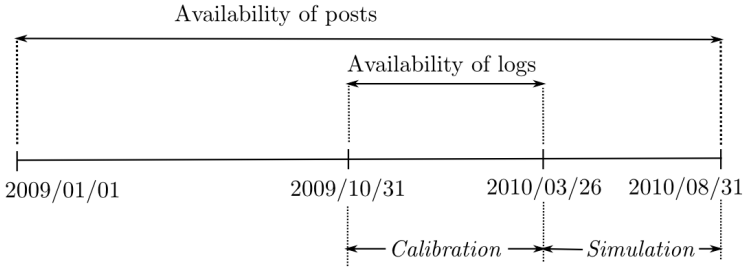


Figure 6.1: Calibration and simulation time windows.

The overall data flow consists of three main phases. In the first phase, the posts are preprocessed and users variables are calibrated. In the second phase, the web forum activity is simulated. Finally, the third phase consists of the analysis of the results produced by the simulation algorithm. The process is illustrated in figure 6.2, where each phase is decomposed into smaller steps. In the present chapter, the methodological framework of the models implementation and evaluation is discussed. In the first place, the experimental data set is described. Subsequently, the data processing is described and the OSN simulation is specified. Finally, the evaluation framework is set.

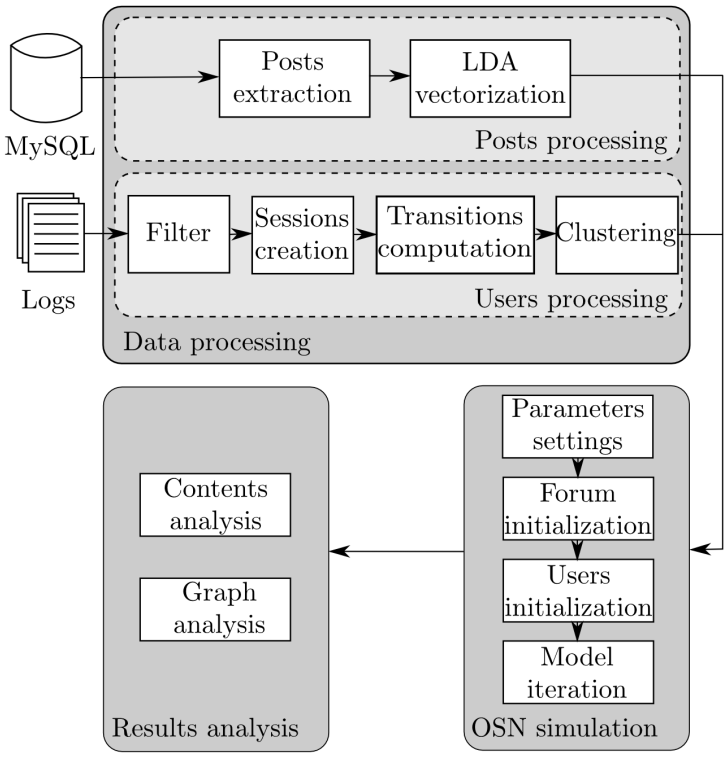


Figure 6.2: Data flow chart.

6.1 Experimental data set

The experimental data set consists of posts and web logs from the OSN Plexilandia, a Chilean web forum. In this section, the Plexilandia OSN is introduced, and the data availability is described.

6.1.1 The Plexilandia web forum

The Plexilandia web forum (`http://www.plexilandia.cl/foro`) was created in 2002. Its name derives from the *plexies*, a famous series of tube amplifiers for electric guitars, fabricated by Marshall [6]. Plexilandia users share their experiences and advices about DIY ("do it yourself") projects, which are mainly amplifiers and effects pedals. As of October of 2012, the forum had 2,715 registered users and 84,147 published messages [11]. The forum activity is organized into six thematic sections: amplifiers, effects, synthesizers, lutherie, professional audio and general discussion (table 6.1).

Thematic section	Subjects	Messages
Amplifiers	2,603	22,213
Effects	3,355	31,198
Synthesizers	33	309
Lutherie	1,324	9,196
Professional audio	179	1,751
General	2,578	19,365

Table 6.1: Thematic sections of Plexilandia and their activity, measured by the number of subjects and published messages. *Source: web site of the forum [11].*

6.1.2 Available data

Two data sets are used in this work. The first one consists of a MySQL database of all posts published between January 2009 and August 2010. The second data set is a collection of five text files which contain the forum's server logs, ranging from November 2009 to March 2010 (table 6.2).

Data set	Number of items	From	To
Posts	10,546	2009/01/01 10:10:13	2010/08/31 23:58:36
Server logs	2,504,440	2009/10/31 07:05:45	2010/03/26 14:53:19

Table 6.2: Experimental data sets.

Among a total of 58 available tables in the MySQL database, three are actually used. The relevant columns are the post id number, the post's topic id number (here, topic means thread; confusion with LDA must be avoided), the poster id number, the post time, the poster IP and the post textual content. Simplified views of these three tables, which only show the columns that are requested by the model, are found in tables 6.3, 6.4 and 6.5.

```
+-----------------+----------------------+------+-----+---------+----------------+
| Field           | Type                 | Null | Key | Default | Extra          |
+-----------------+----------------------+------+-----+---------+----------------+
| post_id         | mediumint(8) unsigned | NO   | PRI | NULL    | auto_increment |
| topic_id        | mediumint(8) unsigned | NO   | MUL | 0       |                |
| poster_id       | mediumint(8)          | NO   | MUL | 0       |                |
| post_time       | int(11)               | NO   | MUL | 0       |                |
| poster_ip       | varchar(8)            | NO   |     |         |                |
+-----------------+----------------------+------+-----+---------+----------------+
```

Table 6.3: Simplified column view of table plxcl_phpbb_posts.

Each server log register contains the requesting IP, the request date, the resourced queried, the HTTP status of the request, the amount of bytes transferred, the web page of origin and the user agent information. A sample of three registers is shown below:

```
186.104.47.249 - - [31/Oct/2009:07:05:45 -0500] "GET /dondecomprar.html
   HTTP/1.1" 200 16072 "http://cl.search.yahoo.com/search?fr=chr-greentree
   _ie&ei=utf-8&type=867034&p=comprar++cosas" "Mozilla/4.0
   (compatible; MSIE 7.0; Windows NT 5.1; GTB6)"
186.104.47.249 - - [31/Oct/2009:07:05:46 -0500] "GET /punto.gif HTTP/1.1"
   200 293 "http://www.plexilandia.cl/dondecomprar.html" "Mozilla/4.0
   (compatible; MSIE 7.0; Windows NT 5.1; GTB6)"
186.104.47.249 - - [31/Oct/2009:07:05:47 -0500] "GET /LOGO.jpg HTTP/1.1"
   200 29293 "http://www.plexilandia.cl/dondecomprar.html" "Mozilla/4.0
   (compatible; MSIE 7.0; Windows NT 5.1; GTB6)"
```

Field	Type	Null	Key	Default	Extra
post_id	mediumint(8) unsigned	NO	PRI	0	
post_text	text	YES		NULL	

Table 6.4: Simplified column view of table plxcl_phpbb_post_texts.

Field	Type	Null	Key	Default	Extra
id	mediumint(8) unsigned	NO	PRI	NULL	auto_increment
user_respuesta_id	mediumint(8)	NO		0	
topic_id	mediumint(8) unsigned	NO		0	
post_id	mediumint(5) unsigned	NO		0	

Table 6.5: Simplified column view of table fav_resumen_posts.

6.2 Data processing

As discussed in chapter 5, the forum-agent system framework is decomposed into the forum structure (section 5.1.1) and the users behavior (section 5.1.2). The forum structure is a simplified representation of the real web forum, consisting of a single list of threads, each of them being in turn a list of posts. Posts are defined by their parent thread, their publication time, their author's id number and their contents. As mentioned in section 4.2.4, the contents are modeled as LDA vectors of topic weights. Therefore, data processing must include a routine for the LDA vectorization of post contents.

On the other hand, the users behavior is defined by a set of user actions and action rules. A central component is the action diagram (figure 5.3) which depicts the set of allowed transitions between actions. Each edge of the diagram requires the sampling of an exponential time with edge-specific rate, and a set of conditional probabilities of transitions given the current node. These rates and conditional probabilities are an input of the model, which are calibrated from the web server logs data. Below, the initialization of both the forum

structure and the action diagram is described.

6.2.1 Posts processing

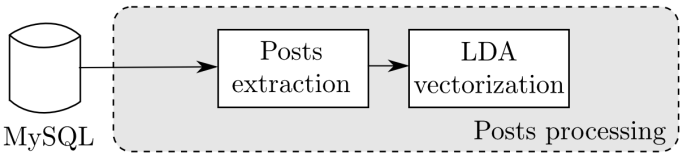


Figure 6.3: Posts processing

Posts processing deals with the initialization of the web forum structure. It is performed by first extracting the posts information from the MySQL database, and then by computing the LDA topic weights vector of each post (figure 6.3). The interface between JAVA and the MySQL database is provided by the MySQL Connector/J library [10]. This allows to perform SQL queries from JAVA code, and to retrieve the result sets within the JAVA virtual machine. The post id number, the parent topic id number, the user id number and the post text are queried from tables `fav_resumen_posts` and `plxcl_phpbb_posts_text`. Then, the JGibbLDA library [8] is invoked for the vectorization of the posts textual contents. In previous work, a LDA vectorization was realized on Plexilandia text contents, with an initial number of 50 topics. The forum's administrators found 33 topics to be relevant, discarding the other 17. Therefore, in this work the number of topics considered is also $k = 33$.

6.2.2 Users processing

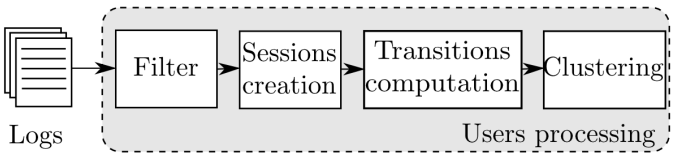


Figure 6.4: Users processing

Users processing deals with the initialization of the users actions diagram. Sessions are built from server logs data and clustered afterwards. Then, users are clustered in function of the previous clustering (figure 6.4). The whole users processing is described in the sequel.

Filter

First, the web logs registers are filtered. Requests from bots and error HTTP status are removed first, and then only URLs related with a thread selection or a message posting are conserved. From a total of $2,504,440$ registers spanning a five month period, $189,692$ are kept after filtering.

Sessions creation

The sessions are created, by first computing preliminary sessions and subsequently inserting missing actions. Preliminary sessions are computed from the filtered registers. These are identified by IP, and consecutive requests of a session must be separated by less than 30 minutes. If they do not, they belong to distinct sessions of the same IP. This way, $47,428$ preliminary sessions are obtained, from which an example is shown below:

```
SESSION ID:23
IP:190.47.183.253
31/Oct/2009:09:59:36 -0500 GET   /foro/viewtopic.php?t=8687
31/Oct/2009:10:00:02 -0500 GET   /foro/viewtopic.php?t=8690
31/Oct/2009:10:03:21 -0500 GET   /foro/viewtopic.php?t=8687
31/Oct/2009:10:04:09 -0500 POST  /foro/posting.php
```

It must be noticed that only thread selections and message postings are specified in preliminary sessions, as in the example above. Final sessions are obtained by adding a `begin-session` action 60 seconds before the first register, and a `end-session` action 60 seconds after the last register. Moreover, `read-a-post` actions are inserted between two consecutive thread selections, and between a thread selection and a consecutive publication. For this purpose, exponential times with mean equal to 30 seconds are iteratively sampled, until the date of the following register is reached. Finally, the user ID of a session is found by matching the log register IP and the posting time with database table `plxcl_phpbb_posts` (6.3). Due to delays between the server logs and the PHPbb system, a difference of 3 seconds in the time matching is allowed. A consequence of the matching is that only sessions with at least one HTTP POST request can be identified with their corresponding user ID numbers. Below lies an example of a final session:

```
SESSION ID =3
POSTER ID  =1
OFF   ON    100107000
ON    T     60000
T     R     38625
R     P     32375
P     R     19254
R     R     6430
R     R     6611
R     R     26929
R     OFF   776
```

A final session therefore contains the list of all the transitions between actions (figure 5.3), with transition times in milliseconds. From the set of 47,428 preliminary sessions, 1,415 final sessions are obtained. The loss is a consequence of the matching process mentioned above.

Transitions computation

The following step is the computation of the transitions involved in each of the 1,415 final sessions. From the action diagram (figure 5.3), it is known that 10 allowed transitions exist. Therefore, each session is characterized in terms of the number of each specific transition that are present, with the corresponding mean time of transition in milliseconds. Tables 6.6 and 6.7 show an example of such characterizations.

Session ID	Poster ID	$n_{ON,T}$	$n_{T,R}$	$n_{R,T}$	$n_{R,R}$	$n_{R,P}$	$n_{R,OFF}$	$n_{P,T}$	$n_{P,R}$	$n_{P,OFF}$	$n_{OFF,ON}$
1	1	1	4	3	3	1	1	0	1	0	0
2	1	1	1	0	13	1	1	0	1	0	1
3	1	1	1	0	3	1	1	0	1	0	1

Table 6.6: Example of sessions characterization by number of transitions.

Session ID	Poster ID	$t_{ON,T}$	$t_{T,R}$	$t_{R,T}$	$t_{R,R}$	$t_{R,P}$	$t_{R,OFF}$	$t_{P,T}$	$t_{P,R}$	$t_{P,OFF}$	$t_{OFF,ON}$
1	1	60,000	12,981	50,094	21,989	13,509	59,567	-	69748	-	-
2	1	60,000	16,768	-	34,828	116,249	6,279	-	28,937	-	10,760,000
3	1	60,000	38,625	-	13,323	32,375	776	-	19,254	-	100,107,000

Table 6.7: Example of sessions characterization by transitions mean times in milliseconds.

Clustering

The final step is the clustering of sessions and users. First, the sessions characterizations are normalized, so each session is represented by a vector in $[0,1]^{20}$. Then, the K-means algorithm with 10 clusters is applied to these vectors, by using the Weka data mining program [9]. The clusters centroids are shown in tables 6.8 and 6.9.

Cluster ID	$n_{ON,T}$	$n_{T,R}$	$n_{R,T}$	$n_{R,R}$	$n_{R,P}$	$n_{R,OFF}$	$n_{P,T}$	$n_{P,R}$	$n_{P,OFF}$	$n_{OFF,ON}$
0	1	1.43	0.39	7.33	1.00	0.89	0.04	0.85	0.11	0.92
1	1	21.50	20.19	72.16	1.41	1.00	0.31	1.09	0.00	0.69
2	1	18.32	16.84	198.32	3.58	0.95	0.47	3.05	0.05	0.58
3	1	9.79	8.74	147.10	1.43	0.98	0.05	1.36	0.02	0.71
4	1	8.59	7.49	27.81	1.00	0.99	0.09	0.89	0.01	0.89
5	1	4.30	3.12	12.00	1.00	0.96	0.18	0.78	0.04	0.89
6	1	7.01	5.96	88.45	1.04	0.99	0.05	0.97	0.01	0.78
7	1	2.17	1.14	31.23	1.00	0.92	0.03	0.90	0.08	0.88
8	1	3.58	2.56	59.89	1.00	0.95	0.02	0.93	0.05	0.90
9	1	4.52	3.25	41.92	2.16	0.90	0.26	1.80	0.10	0.93

Table 6.8: Sessions clusters centroids, transitions frequencies.

Cluster ID	$t_{ON,T}$	$t_{T,R}$	$t_{R,T}$	$t_{R,R}$	$t_{R,P}$	$t_{R,OFF}$	$t_{P,T}$	$t_{P,R}$	$t_{P,OFF}$
0	60,000	25,439	4,817	23,061	27,297	18,182	1,038	17,583	6,786
1	60,000	15,324	16,365	26,336	32,305	20,853	9,063	20,576	0
2	60,000	21,913	21,657	27,718	23,816	16,433	8,737	30,537	3,158
3	60,000	23,120	23,204	28,115	26,586	23,595	6,90	28,231	1,429
4	60,000	16,011	14,864	24,524	22,169	21,890	2,388	17,518	706
5	60,000	17,112	16,158	22,999	26,546	22,541	4,882	14,647	2,258
6	60,000	23,819	20,801	27,733	30,654	22,525	1,425	27,015	822
7	60,000	25,680	14,512	27,627	27,453	19,416	1,119	20,473	4,541
8	60,000	26,159	21,455	28,255	24,288	19,329	395	25,285	3,158
9	60,000	24,382	21,776	25,811	28,516	17,933	6,652	21,800	5,902

Table 6.9: Sessions clusters centroids, transitions mean times in milliseconds.

The sessions having been clustered, a second clustering is then applied to users. Each of them is characterized in a space that specifies the proportions of each session type in his own history, plus the mean time between two sessions. The clustering is performed again through K-means, with 2 clusters. The centroids obtained are shown in table 6.10, where $\%s_{\mathrm{i}}$ denotes the percentage of sessions of type i (defined by the tables 6.8 and 6.9). It is observed that the first cluster of users is heavily weighted on sessions of type 0, which are short sessions with few transitions. It therefore represents users that visit the web forum occasionally. In contrast, the second cluster of users is much more balanced across session types.

Cluster ID	$\%s_0$	$\%s_1$	$\%s_2$	$\%s_3$	$\%s_4$	$\%s_5$	$\%s_6$	$\%s_7$	$\%s_8$	$\%s_9$	$t_{OFF,ON}$
0	68.76%	1.83%	0.06%	0.30%	2.47%	5.22%	2.51%	9.87%	3.79%	5.20%	715,748,239
1	13.10%	6.72%	4.27%	6.89%	8.43%	16.24%	10.12%	14.73%	9.79%	9.71%	1,062,213,507

Table 6.10: Users clusters centroids.

6.3 OSN simulation

In the data processing phase, the posts contents are transformed to LDA vectors, and the users are clustered according to their sessions behavior. Ten clusters of sessions are computed, so each cluster has its own parameters for the action diagram. In the OSN simulation

phase, four steps are executed (figure 6.5). First, the model parameters are set. The forum structure is then initialized, and so are the users. Finally, the model is run.

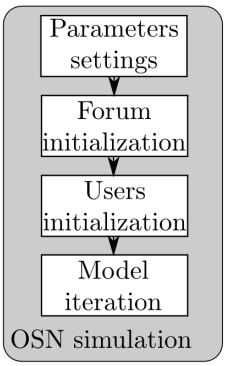


Figure 6.5: OSN simulation

Parameters settings

First, the model is chosen among the FreqLCA, FreqMax, FreqLogit and FreqRandom models. Then, the parameters associated to action rules given by equations 5.3, 5.4 and 5.5 are set. Namely, $(c_\pi, c_\mu, c_\nu) \in \mathcal{P} = \{0, 0.25, 0.5, 0.75, 1\}^3$. At this stage, the FreqRandom is completely defined, as its `choose-a-thread` and `choose-a-post` action rules do not depend on any parameter (section 5.2.3). However, the FreqLCA, FreqMax and FreqLogit models rely on the computation of logit probabilities, which depend on three parameters: α_t, α_μ and α_ν. These are arbitrarily set to $\alpha_t = \alpha_\mu = \alpha_\nu = 1$. Now, the FreqMax (section5.2.4) and FreqLogit (section 5.2.2) are completely defined. Only the FreqLCA (section 5.2.1) remains, as parameters associated to the LCA model need to be set. The decay parameter is set to $k = 10$, and the inhibition parameter is set to $w = 10$. Finally, the threshold Z is set to $Z = 1$ (table 6.11).

Parameter	c_π	c_μ	c_ν	α_μ	α_ν	α_t	k	w	Z
FreqLCA	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	1	1	1	10	10	1
FreqMax	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	1	1	1	-	-	-
FreqLogit	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	1	1	1	-	-	-
FreqRandom	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	$\in \{0, 0.25, 0.5, 0.75, 1\}$	-	-	-	-	-	-

Table 6.11: Parameters settings.

Forum initialization

In the second step of the OSN simulation phase, the forum structure is loaded as it was at 2010/03/26 : 14 : 53 : 19.

Users initialization

In the third step of the OSN simulation phase, the users are initialized by loading the clusters found in the data processing phase. First, each user is assigned to its corresponding cluster. Then, it is initialized in the OFF state of the action diagram. Subsequently, the next action is chosen at random. As the user is at OFF, the only possibility is ON (`begin-session`), according to the action diagram (figure 5.3). Finally, an exponential time is sampled for the transition from OFF to ON, according to the rate given by the user's cluster (table 6.10). Let u be the user, t_0 the simulation initial time and Δt the exponential time sampled. Then the action $(u, t_0 + \Delta t, ON)$ is now completely defined: the user u will perform the action ON at instant $t_0 + \Delta t$. This action is added to a schedule, that sorts actions by increasing execution time. The overall step of users initialization is summarized by the following pseudo algorithm:

initialize simulator time at $t_0 = 2010/03/26 : 14 : 53 : 19$
for u in users
  assign u to its cluster $c(u) \in \{0, 1\}$
  initialize current state at OFF
  choose next action: ON
  sample exponential time t of transition from OFF to ON according to $c(u)$
  add action $(u, t_0 + \Delta t, ON)$ to schedule
endfor

Model iteration

In the final step of the OSN simulation phase, the initial actions are executed and new actions are added to schedule. In each iteration of the schedule, the next action (u, t, A) is removed, where t is minimal and $A \in \{ON, T, R, P, OFF\}$. The action A is executed by user u at instant t, and then the next action A' is chosen at random from the decision set $\delta^+(A)$. According to the action diagram (figure 5.3) there are five cases:

- if $A = ON$, then $\delta^+(A) = \{T\}$
- if $A = T$, then $\delta^+(A) = \{R\}$
- if $A = R$, then $\delta^+(A) = \{T, R, P, OFF\}$
- if $A = P$, then $\delta^+(A) = \{T, R, OFF\}$
- if $A = OFF$, then $\delta^+(A) = \{ON\}$

If the decision set contains more than one action, then the next action is chosen at random depending on the current session type of the user and the transitions frequencies (table 6.8). If $A \neq OFF$, then the transition time Δt is an exponential time sampled with mean given by table 6.9, depending on the current session type. If $A = OFF$, then it is sampled with mean given by table 6.10, depending on the user cluster. If the next action date $t + \Delta t$ is

sooner than the simulation end time, the action $(u, t + \Delta t, A')$ is added to schedule. On the contrary, it is discarded and A was the last action executed by user u during the simulation period. The model iteration is summarized as follows:

while schedule is not empty
  remove action (u, t, A) with lowest execution time t
  the user u performs action A at t
  choose next action A' and sample transition time Δt
  if $t + \Delta t \leq 2010/08/31 : 23 : 58 : 36$
    add action $(u, t + \Delta t, A')$ to schedule
  endif
endwhile

6.4 Results analysis

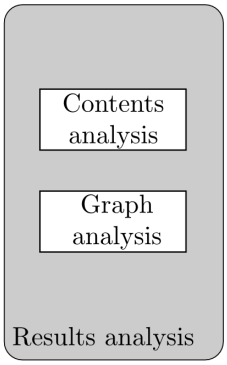


Figure 6.6: Results analysis.

The models performance is assessed in function of the two hypotheses of this work, which are:

R.H. 1 (*superiority of LCA as the decision mechanism*) the LCA model, as the underlying mechanism of decision, gives the best prediction of actual contents and graph generation, in comparison with a voter model, a Logit-based model and a deterministic model.

R.H. 2 (*decay of contents variance over time*) the variance of the LDA profiles across users decreases with time.

Three dimensions of analysis are involved in the evaluation of the models performance. The first one concerns the similarity between real and simulated contents, characterized as LDA vectors. For this purpose, the mean average percentual error (MAPE) is used as the performance measure. The second one is the distance between the real and the simulated

interaction graphs, generated by the users' activity in the forum. The generated graph structure is analysed by the means of a F-measure. Finally, the third dimension of analysis is focused on the evolution of the variance of contents over time, where it is expected to find a negative trend. For this purpose, an ordinary least squares (OLS) estimation of the trend is performed. The contents analysis step includes contents generation and contents variance, while the graph analysis step includes graph generation (figure 6.6). The previous dimensions of analysis are studied on various time scales, ranging from the week to the global time window of analysis. When a time scale is deemed too wide for a particular dimension of analysis, it is not considered (table 6.12).

			Time scale		
Dimension of analysis	Related hypothesis	Performance indicator	Weekly	Monthly	Global
Contents generation	R.H. 1	MAPE	Yes	Yes	No
Graph generation	R.H. 1	F-measure	Yes	Yes	Yes
Contents variance	R.H. 2	OLS estimator	Yes	No	No

Table 6.12: Summary of the dimensions of analysis, performance indicators and time scales.

In the following, each of these analysis dimensions, and their performance indicators, are discussed.

6.4.1 Contents generation analysis

The evaluation of the similarity of simulated contents, with respect to real information, is aimed at certifying that the model captures correctly the creation of text. As the main subject of this work is the diffusion of information, it is therefore an important dimension that must be analysed.

Performance indicator

Text contents are treated as fixed-length LDA vectors. The question to be answered is: how much the components of the LDA vectors (both simulated and real) differ, on average? The MAPE is proposed as the performance indicator, which delivers the average absolute value of the relative differences between actual and predicted components. Let $A \in \mathbb{R}^n$ be a vector of n actual values, and $P \in \mathbb{R}^n$ its predicted — or estimated— equivalent. The MAPE between A and P is defined as:

$$\mathrm{MAPE}(A, P) = \frac{1}{n} \sum_{\mathrm{i}=1}^{n} \left| \frac{A_{\mathrm{i}} - P_{\mathrm{i}}}{A_{\mathrm{i}}} \right| \tag{6.1}$$

It must be noted, on the one hand, that differences of components always increment the MAPE because of the absolute value, and, on the other hand, that the MAPE comes in percentual units (see figure 6.7 for an example).

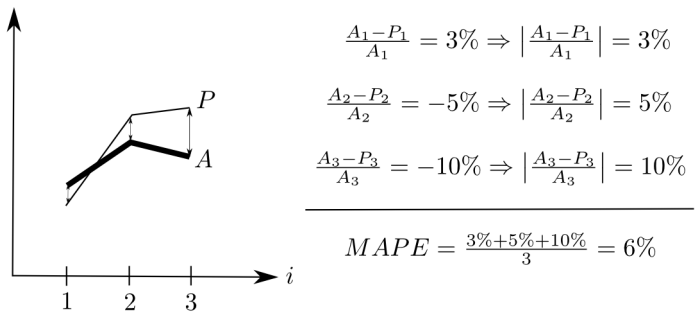


Figure 6.7: Example of a MAPE calculation.

Evaluation procedure

The case of a weekly time scale will be treated below, the other cases being analogous. Suppose that the simulation period is made of T weeks. First, for each week $t \in \{1, \ldots, T\}$, the mean LDA vectors, related to the publications generated during the week, are calculated. If r_t real posts and s_t simulated ones have been generated during week t, then these mean vectors are:

$$\mathbf{p}^t_{REAL} = \frac{1}{r_t} \sum_{\mathrm{i}=1}^{r_t} \mathbf{p}^{t,\mathrm{i}}_{REAL} \tag{6.2}$$

$$\mathbf{p}^t_{SIM} = \frac{1}{s_t} \sum_{\mathrm{i}=1}^{s_t} \mathbf{p}^{t,\mathrm{i}}_{SIM} \tag{6.3}$$

The MAPE of week t is then computed:

$$\mathrm{MAPE}_t = \mathrm{MAPE}(\mathbf{p}^t_{REAL}, \mathbf{p}^t_{SIM}) \tag{6.4}$$

Finally, in order to obtain a single performance indicator, the errors are averaged over time:

$$\overline{\mathrm{MAPE}} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{MAPE}_t \tag{6.5}$$

6.4.2 Graph generation analysis

As the diffusion of information occurs over a social network, the accuracy of the predicted graphs is relevant too. The evaluation of the similarity between real and simulated graphs is considered in the following.

Performance indicator

Recall and precision are two performance indicators of a predictive model whose dependent variable is binary. Consider a set of N items where each of them can be either of type A or

B. Further, A is the reference type, and its members are called *positives*, while the members of B are called *negatives*. With a predictive model that assigns a predicted type for each item, four possibilities may arise:

the item actual type is A and its predicted type is A: this corresponds to a correct match by the predictive model. The model assignment is a *true positive*.

the item actual type is A and its predicted type is B: while the item is actually a positive, the model labels it as a negative. This model assignment is called a *false negative*.

the item actual type is B and its predicted type is B: the model assigns correctly the item to the B type, and so this is a *true negative*.

the item actual type is B and its predicted type is A: the model incorrectly assigns the negative item to the category of positives. The model assignment is a *false positive*.

These cases may be depicted graphically, as shown in figure 6.8.

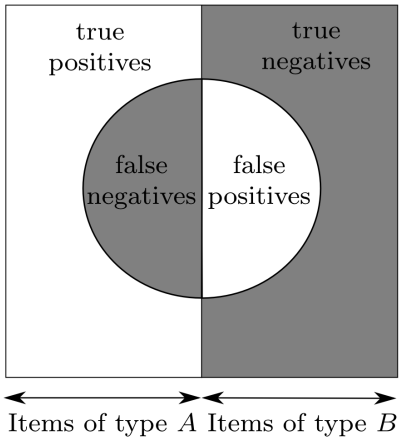


Figure 6.8: From a total of N items, N_A are actually of type A and N_B are actually of type B. The grey shaded area depicts the items which have been assigned by a predictive model to the B category; the white area represents the items labeled as A.

Let tp be the number of true positives, fn the number of false negatives, tn the number of true negatives and fp the number of false positives. Then, from the previous discussion it follows that:

$$N = tp + fn + tn + tp \tag{6.6}$$

$$N_A = tp + fn \tag{6.7}$$

$$N_B = tn + fp \tag{6.8}$$

In this context, two measures of the model's predictive performance are its *precision* and *recall*. The precision is defined as the fraction of correctly identified positive items over all items identified as positive:

$$precision = \frac{tp}{tp + fp} \tag{6.9}$$

The recall is the fraction of correctly identified positive items over all items that are actually positive:

$$recall = \frac{tp}{tp + fn} \tag{6.10}$$

Hence, there is trade-off between precision and recall. Indeed, a predictive model that labels all items as positive, reaches a maximal recall of 1.0. But its precision would, very likely, be small, as all negative items are wrongly labeled as positives. On the contrary, a model that (correctly) assigns a unique item to the A category produces a maximal precision of 1.0, but with a very low recall. Therefore, in order to obtain a single measure of predictive performance, the F-measure is introduced as follows:

$$F_{measure} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6.11}$$

Evaluation procedure

The model simulates forum activity from which an interaction graph may be built. The interaction graph is modeled as a directed graph $G = (V, E)$, where V is the set of users and $E = \{(u, v) : u, v \in V\}$ is the set of edges between them. The existence of an edge $(u, v) \in E$ means that v has replied to u in some way. Three network topologies are considered here: all-previous, creator and last-post. These topologies produce different graphs for the same forum structure. For example, consider a thread configuration as depicted in figure 6.9.

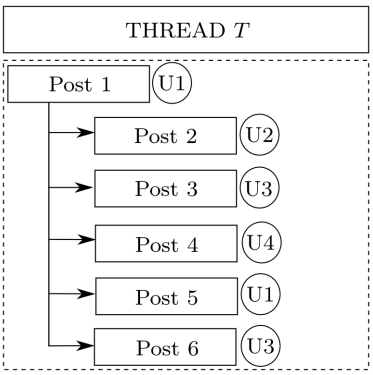


Figure 6.9: Example of thread configuration. Circles show users.

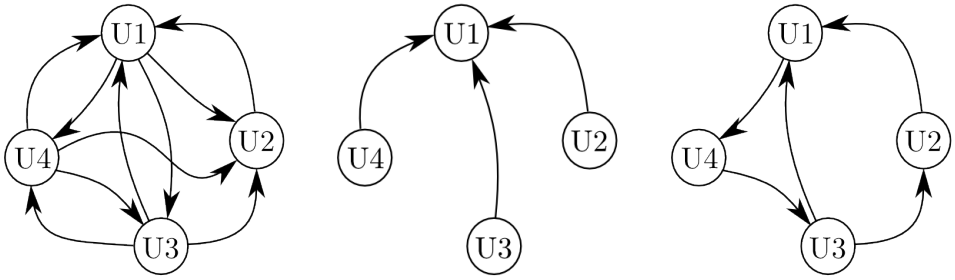
The network topologies are built as follows:

all-previous network when a user publishes a post, an edge is drawn from the user to each of the previous posts authors

creator network when a user publishes a post, an edge is drawn from the user to the creator of the thread

last-post network when a user publishes a post, an edge is drawn from the user to the last post author

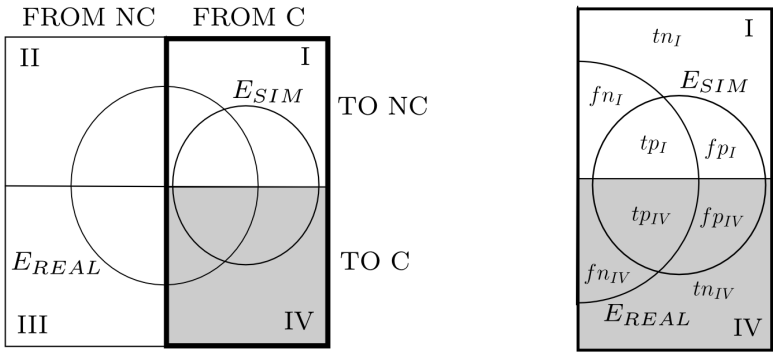
Figure 6.10 shows the three networks resulting from the example above. The graph at OSN level is obtained as the union of graphs computed at thread level.



(a) All-previous network. (b) Creator network. (c) Last post network.

Figure 6.10: Network topologies.

A distinction is made between the set C of users who have been calibrated, and the set NC of users who have not ($V = C \cup NC$). As the simulation considers the activity of calibrated users, only edges of the kind $(u, v), u \in C, v \in V$ are produced. Therefore, the precision-recall analysis is possible only over this subset of the global set of possible edges E. Moreover, the focus of interest is placed on the performance of the models for the calibrated users. Hence the precision-recall analysis is performed on edges $(u, v), u, v \in C$ (figure 6.11). The F-measure is obtained by using equations 6.9, 6.10 and 6.11 to quadrant IV.



(a) The simulation domain is surrounded by a thick line.

(b) Detail of the simulation domain.

Figure 6.11: The shaded area depicts the quadrant where precision-recall is actually performed. Subfigure **(a)**: the edges (u, v) are classified, depending on if the origin and the destination users are calibrated or not. $E_{REAL} \subset E$ is the subset of actual edges, and $E_{SIM} \subset E$ is the subset of edges predicted by simulation. As the activity of calibrated users is the only one to be simulated, then predicted edges necessarily belong to the quadrants I and IV. Subfigure **(b)**: classification of edges by quadrant, showing true positives, true negatives, false positives and false negatives.

6.4.3 Contents variance temporal analysis

At the beginning of the simulation, the forum's users have heterogeneous preferences, and therefore post differently one from another. But as time passes, since no external input of information is provided during the simulation, the system should converge to an equilibrium. The contents variance is the dimension of analysis which is therefore related to the diffusion of information among the users.

Performance indicator

A linear model for the dependent variable y, in function of the independent variables $x_1, \ldots, x_m$ is:

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_{\mathrm{i},1} + \ldots + \beta_m x_{\mathrm{i},m} \quad \mathrm{i} = 1, \ldots, n \tag{6.12}$$

where y_{i} is the i^{th} observation of variable y, and $x_{\mathrm{i},j}$ is the i^{th} observation of variable x_j. In matrix form:

$$Y = X\beta \tag{6.13}$$

where Y is a $n \times 1$ matrix, X is a $n \times (1+m)$ matrix and β is a $(1+m) \times 1$ matrix. The sum of squared residuals (SSR) is equal to $(Y - X\beta)^T(Y - X\beta)$, which is minimized over all possible β as follows:

$$\min_{\beta} SSR(\beta) = (Y - X\beta)^T(Y - X\beta) \tag{6.14}$$

$$= Y^TY + \beta^T X^T X\beta - 2Y^T X\beta \tag{6.15}$$

$$\Rightarrow \frac{\mathrm{d}SSR}{\mathrm{d}\beta} = 0 \tag{6.16}$$

$$\Rightarrow 2X^TX\hat{\beta} - 2X^TY = 0 \tag{6.17}$$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^TY \tag{6.18}$$

where $\hat{\beta}$ is the OLS estimator of β. For the case $m = 1$, the simple linear model $y_{\mathrm{i}} = a + bx_{\mathrm{i}}$ is obtained, and the ordinary least squares estimator is:

$$\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}, \; \hat{b} = \frac{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})(y_{\mathrm{i}} - \bar{y})}{\sum_{\mathrm{i}=1}^{n}(x_{\mathrm{i}} - \bar{x})(x_{\mathrm{i}} - \bar{x})}, \; \hat{a} = \bar{y} - \hat{b}\bar{x} \tag{6.19}$$

Evaluation procedure

For each week t, the vectors of topic components standard deviations, $\vec{\sigma}^t_{REAL}$ and $\vec{\sigma}^t_{SIM}$ are computed. Then, their components are in turn averaged, thus obtaining the mean standard deviation:

$$\sigma^t_{REAL} = \frac{\sum_{\mathrm{i}=1}^{N_\theta}(\vec{\sigma}^t_{REAL})_{\mathrm{i}}}{N_\theta} \tag{6.20}$$

$$\sigma^t_{SIM} = \frac{\sum_{\mathrm{i}=1}^{N_\theta}(\vec{\sigma}^t_{SIM})_{\mathrm{i}}}{N_\theta} \tag{6.21}$$

In order to detect a possible convergence (or divergence) in the forum's discussion over time, a linear function of time is fitted — with OLS — to the time series σ^t_{REAL} and σ^t_{SIM}:

$$\sigma^t_{REAL} \approx \hat{a}_{REAL} + \hat{b}_{REAL} \cdot t \tag{6.22}$$

$$\sigma^t_{SIM} \approx \hat{a}_{SIM} + \hat{b}_{SIM} \cdot t \tag{6.23}$$

Of particular interest is the value of $\hat{b}$ trend coefficients, which reveals the variance trend over time. By equation 6.19:

$$\hat{b}_{REAL} = \frac{\sum_{t=1}^{T}(t - \frac{T+1}{2})(\sigma^t_{REAL} - \overline{\sigma_{REAL}})}{\sum_{t=1}^{T}(t - \frac{T+1}{2})^2} \tag{6.24}$$

$$\hat{b}_{SIM} = \frac{\sum_{t=1}^{T}(t - \frac{T+1}{2})(\sigma^t_{SIM} - \overline{\sigma_{SIM}})}{\sum_{t=1}^{T}(t - \frac{T+1}{2})^2} \tag{6.25}$$

6.5 Discussion

A complete framework for data processing, simulation and analysis is designed and implemented. Textual contents are represented as LDA vectors of topic weights and users are clustered. There is a striking similarity between the user clustering process and the LDA generative process. Indeed, the user cluster type can be seen as a mixture of multinomial distributions. The same applies to session clusters. Therefore, the process described in this chapter is analogous to a generative process where the user cluster is first sampled, followed by the session cluster. This suggests the application of Bayesian networks in the processing of web logs, with the advantage of conceptual unification.

The models FreqLCA, FreqMax, FreqRandom and FreqLogit are tested on the $\mathcal{P} = \{0, 0.25, 0.5, 0.75, 1\}^3$ parameters grid for (c_π, c_μ, c_ν). The performance of the models is measured in terms of contents generation, graph generation and temporal contents variance. Regarding graph generation, three network topologies are considered: all-previous, creator and last-post. These topologies measure different aspects. The all-previous graph is concerned with general interaction. The creator graph measures if the users post in the right threads. The last-post graph measures if the sequence of simulated posts is correct within a thread. The lowest density is expected for the creator graph, followed by last-post. The all-previous network is expected to exhibit the highest density. It may be therefore inferred that the all-previous graph is the easiest to reproduce, while creator and last-post topologies demand greater simulation precision.

Chapter 7

Results

In the methodological framework, three dimensions of analysis have been defined : contents generation, graph generation and contents variance. The objective of contents and graph generation analysis is to prove that the main model produces a more realistic behavior of the forum than the benchmark models explained in Chapter 5. The contents variance is related to the more general problem of information diffusion among users. As discussed in section 5.1.2, three parameters are involved in the action rules of users: a preference parameter c_μ, a publication parameter c_π and a social image parameter c_ν. Each of the four models (FreqLCA, FreqMax, FreqLogit, FreqRandom) is tested on a 125-points parameters grid $(c_\pi, c_\mu, c_\nu) \in \mathcal{P} = \{0, 0.25, 0.5, 0.75, 1\}^3$, with 300 iterations on each point. Three parameters — specific to the models FreqLCA, FreqLogit and FreqMax — rule the valuation of text similarity, recency and image compatibility at the decision level: α_μ,α_t and α_ν. These have been assumed to be all equal to one (section 6.3).

In the present chapter, the results obtained for the implemented models on the grid points are presented. First, the weekly and monthly MAPE of the simulated models, versus the real posts generated during the simulation window, are discussed. Then, the F-measures are shown, for graphs generated weekly, monthly and globally, according to three graph topologies: all previous, creator and last post. Finally, the OLS estimators of the contents variance temporal trend are considered.

7.1 Contents generation

The performance indicator for the analysis of contents generation is the mean average percentual error (MAPE; see Chapter 6, section 6.4.1). The MAPE measures the average percentual difference between the simulated topic profiles and the real ones. Two time scales are considered. For the first one, the mean topic profiles are computed on a weekly basis and their MAPE is computed. Then, the average MAPE over weeks is calculated. The process is similar for the second time scale, in which the MAPE are computed on a monthly basis. In

each case, the results are compared with a reference MAPE, which corresponds to a flat topic profile. The reference MAPE would be obtained by a model where the users systematically publish flat topic vectors. In the sequel, the average weekly MAPE are analysed, then the average monthly MAPE are discussed. Complete results are found in appendix C.

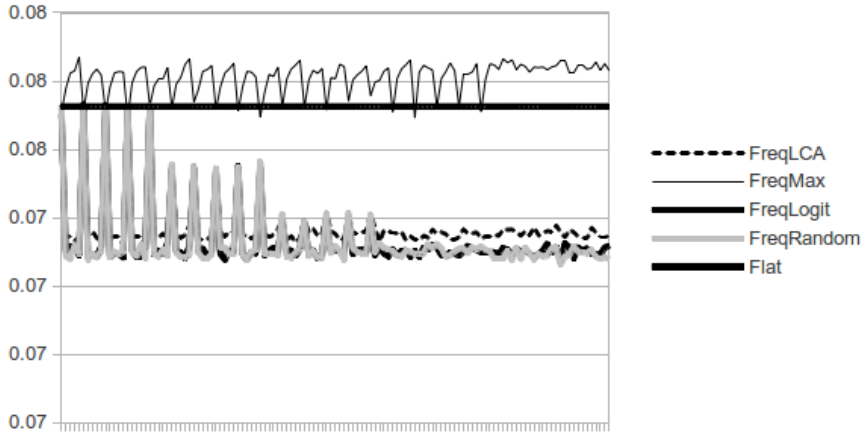
7.1.1 Average weekly MAPE

A reference MAPE of 7.724% is obtained. The main model FreqLCA is consistently outperformed by the random model FreqRandom and the logit model FreqLogit (table 7.1; figure 7.1). The random model FreqRandom ranks first with a best MAPE value of 7.263%. It is followed closely by the logit model FreqLogit in the second position, which best MAPE value is 7.272. The main model ranks third (7.321%), and the worst model is the deterministic model FreqMax (7.693%).

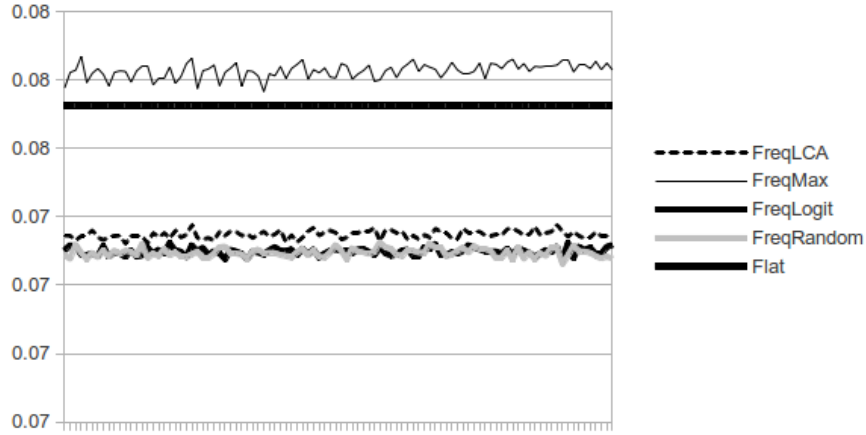
No consistent pattern in the parameters values is observed, except for the deterministic model FreqMax where the best values are obtained for $c_\mu = 0$. From equation 5.3 (section 5.1.2), this means that the posts read by the users have no influence on their beliefs: μ is constant over time. Interestingly, the condition $c_\mu = 0$ is consistently observed over the worst MAPE results for the other models (appendix C).

Model	Rank	c_π	c_μ	c_ν	MAPE	Information gain
FreqLCA	1	0	1	0.5	7.321%	0.403%
FreqLCA	2	0	1	0.75	7.323%	0.402%
FreqLCA	3	0.5	0.25	0	7.326%	0.399%
FreqLCA	4	0.5	0.75	0	7.327%	0.398%
FreqLCA	5	0.5	0.5	1	7.329%	0.396%
FreqMax	1	0.75	0	0.25	7.693%	0.032%
FreqMax	2	0	0	0.75	7.693%	0.031%
FreqMax	3	0.25	0	1	7.695%	0.030%
FreqMax	4	0.75	0	0	7.709%	0.015%
FreqMax	5	0.75	0	1	7.710%	0.014%
FreqLogit	1	1	0.75	0.5	7.272%	0.452%
FreqLogit	2	0.25	0.5	0.75	7.275%	0.449%
FreqLogit	3	0.25	0.5	0.5	7.276%	0.448%
FreqLogit	4	1	0.25	0.75	7.281%	0.444%
FreqLogit	5	0.5	0.75	0.25	7.281%	0.444%
FreqRandom	1	1	0.75	0.5	7.263%	0.462%
FreqRandom	2	1	0.5	0	7.274%	0.451%
FreqRandom	3	1	0.5	0.25	7.276%	0.449%
FreqRandom	4	0.5	1	0.5	7.276%	0.449%
FreqRandom	5	0.25	0.5	0.75	7.277%	0.448%

Table 7.1: Best average weekly MAPE. For each of the four models, the five best combinations of parameters are shown, ordered by increasing MAPE. The information gain is defined as the reference MAPE minus the model MAPE.



(a) Average weekly MAPE ($n = 125$). The upward peaks for FreqLCA, FreqLogit and FreqRandom occur for $c_\mu = 0$, which corresponds to downward peaks of FreqMax. Note that the upward peaks decrease when c_π increases, and are indifferent with respect to c_ν.



(b) Average weekly MAPE, excluding tuples where $c_\mu = 0$ ($n = 100$). The peaks disappear, although a periodicity is still observed, in particular for FreqMax. The obtained MAPE is relatively insensitive to changes in the parameters values.

Figure 7.1: Average weekly MAPE versus the tuples in $\mathcal{P}$. The parameters c_μ, c_ν and c_π are ordered in increasing order. The parameter c_μ has cycles of length 5, c_ν of length 25 and c_π of length 125. For example, in the ten first points of the x-axis, the c_μ values are 0/0.25/0.5/0.75/1/0/0.25/0.5/0.75/1; the c_ν values are 0/0/0/0/0/0.25/0.25/0.25/0.25/0.25; the c_π values are 0/0/0/0/0/0/0/0/0/0.

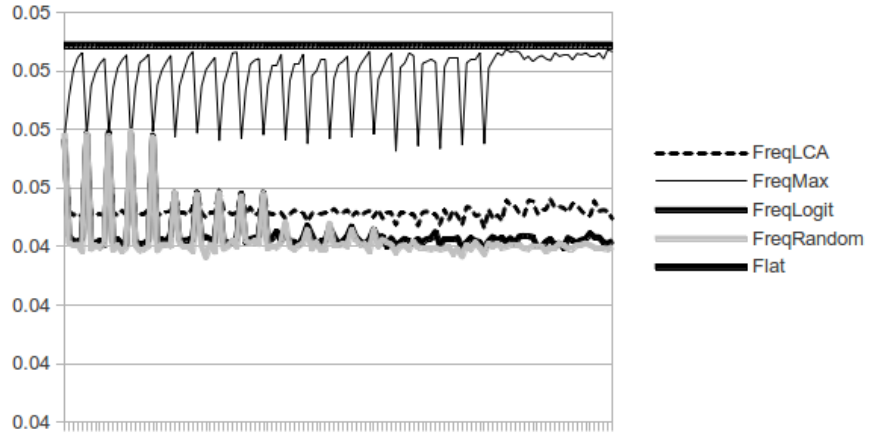
7.1.2 Average monthly MAPE

Better MAPE values are obtained throughout the models, and the reference monthly MAPE improves to 5.089%. A plausible explanation is the averaging of errors, as more posts are considered for the MAPE calculations. The main model FreqLCA is outperformed by the random model FreqRandom and the logit model FreqLogit (table 7.2; figure 7.2). The random model FreqRandom ranks first with a best MAPE value of 4.362%. The logit model ranks second with a best MAPE value of 4.392%. The main model FreqLCA ranks third with a best MAPE value of 4.463%. Finally, the deterministic model FreqMax ranks fourth with a best MAPE value of 4.725%.

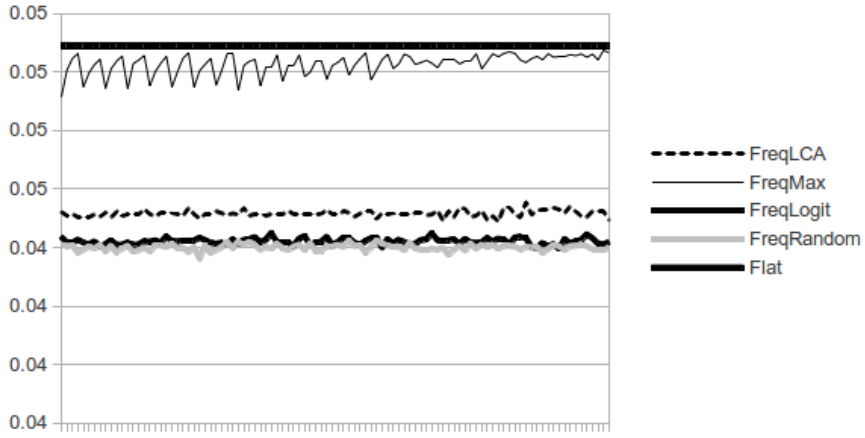
Much more regular results are obtained for the parameters values that yield the best MAPE values. Interestingly, the main model FreqLCA and the deterministic model FreqMax exhibit a very similar behavior, with $c_\mu = 0$ and $c_\pi = 0.75$. The meaning of c_μ has already been discussed above. According to equation 5.4, the c_π value of 0.75 implies that a new post is equal to the weighted average of the last post read (75%) and the user's preference (25%). This corresponds to a profile of users whose preferences are fixed ($c_\mu = 0$), but whose publishing behavior is very dependent of the other users' publications ($c_\pi = 0.75$). The logit model FreqLogit and the random model FreqRandom also show high values of c_π, but no regularity is observed for c_μ. Throughout the models, no regularity is observed for parameter c_ν. The value $c_\mu = 0$ is characteristic of the worst quality of fit for all models, except the deterministic model FreqMax (see appendix C for more details).

Model	Rank	c_π	c_μ	c_ν	MAPE	Information gain
FreqLCA	1	0.75	0	1	4.463%	0.626%
FreqLCA	2	0.75	0	0.25	4.470%	0.618%
FreqLCA	3	0.75	0	0	4.473%	0.616%
FreqLCA	4	0.75	0	0.5	4.474%	0.615%
FreqLCA	5	0.75	0	0.75	4.474%	0.614%
FreqMax	1	0.75	0	0	4.725%	0.363%
FreqMax	2	0.75	0	0.5	4.734%	0.354%
FreqMax	3	0.75	0	0.25	4.744%	0.345%
FreqMax	4	0.75	0	0.75	4.747%	0.342%
FreqMax	5	0.75	0	1	4.751%	0.338%
FreqLogit	1	0.75	0	1	4.392%	0.697%
FreqLogit	2	1	0.75	0.5	4.395%	0.693%
FreqLogit	3	0.75	0	0.75	4.395%	0.693%
FreqLogit	4	1	0.75	0.25	4.399%	0.690%
FreqLogit	5	0.5	0.75	1	4.399%	0.689%
FreqRandom	1	0.25	0.5	0.25	4.362%	0.726%
FreqRandom	2	0.75	0	0.75	4.365%	0.723%
FreqRandom	3	0.75	0	1	4.374%	0.714%
FreqRandom	4	0.75	0.75	0.5	4.375%	0.713%
FreqRandom	5	0.75	0	0	4.379%	0.710%

Table 7.2: Best monthly MAPE and information gain with respect to the reference monthly MAPE.



(a) Average monthly MAPE ($n = 125$).



(b) Average monthly MAPE, excluding tuples where $c_\mu = 0$ ($n = 100$).

Figure 7.2: Average monthly MAPE versus the tuples in $\mathcal{P}$.

7.2 Graph generation

The performance indicator for graph generation is the F-measure, as discussed in section 6.4.2. The F-measure is the harmonic mean of precision and recall, and therefore constitutes a single indicator of the simulated graph quality. Three time scales are considered: the week, the month and the whole period of simulation. In the first case, simulated graphs are built on a weekly basis, and the F-measure with respect to the real graph is computed and averaged over weeks. The other cases are similar. For each time scale, the three graph topologies are analysed : all previous, creator and last post. Three graphs are therefore built per time unit. Below, the results obtained are presented by time scale. Complete results are found in appendix D.

7.2.1 Average weekly F-measure

Overall, very low F-measure values are obtained across graph topologies and models (table 7.3). Better values are observed for the all previous topology, due to the greater density of the graph. Moreover, for that topology, the FreqMax model exhibits clearly a better performance than other models, with a best F-measure of 5.577%. No significant difference is observed between the other models. For the creator topology, the F-measure of the FreqMax model is undefined because of very low recall and precision levels. For the last post topology, the FreqMax performs slightly better than other models. Interestingly, the worst results for the FreqMax model are obtained for $c_\nu = 1$ (see appendix D). By equation 5.5, it means that the social image is constantly updated to the last post published by the user.

7.2.2 Average monthly F-measure

An overall improvement in the value of F-measure is measured in the monthly case. Though, they are still very low, and only the FreqMax model manages to pass the 10% level with the all previous topology. The FreqMax is clearly the best model for the all previous ($F = 12.434\%$) and last post (2.065%) topologies, but is slightly surpassed by the FreqLCA and the FreqRandom *ex aequo* for the creator topology ($F = 1.764\%$). No clear regularities are observed concerning the parameters.

7.2.3 Average global F-measure

The overall improvement of F-measure values further accentuates when graphs are built over the whole period of simulation. The FreqLCA, FreqLogit and FreqRandom reach levels above 20% for the all previous topology. Surprisingly, the FreqMax is the worst model for

		All previous				Creator				Last post			
Model	Rank	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F
FreqLCA	1	0.75	0.75	0.75	2.722%	0	0.5	0.75	0.508%	0	1	0.25	0.519%
FreqLCA	2	0.75	0	0.5	2.722%	0.75	0.75	0.75	0.508%	0	0.25	0.25	0.513%
FreqLCA	3	0.5	0.75	0.25	2.720%	0	0	0.75	0.508%	1	1	1	0.509%
FreqLCA	4	0.75	0.25	0	2.718%	0	1	0.25	0.508%	0.5	1	1	0.508%
FreqLCA	5	0	0.5	0	2.714%	1	0.75	1	0.505%	0.75	0.5	0	0.508%
FreqMax	1	0.75	0.25	0.25	5.577%	-	-	-	-	0	0.75	0	0.669%
FreqMax	2	1	0.75	0.25	5.571%	-	-	-	-	0	1	0	0.666%
FreqMax	3	0	0.75	0.5	5.550%	-	-	-	-	0.25	0	0.5	0.659%
FreqMax	4	0.5	0	0	5.536%	-	-	-	-	1	0	0.5	0.658%
FreqMax	5	1	0.25	0	5.535%	-	-	-	-	0.5	0	0	0.656%
FreqLogit	1	1	1	1	2.764%	0.75	1	0.5	0.515%	0.75	0	0.25	0.530%
FreqLogit	2	0.75	0	0.75	2.746%	0.25	0.5	0.25	0.515%	0	0.5	0	0.529%
FreqLogit	3	0.25	1	1	2.742%	0.75	0	0.75	0.515%	0.75	1	0.25	0.528%
FreqLogit	4	0.5	0	0	2.737%	1	0.25	0	0.513%	0	1	1	0.524%
FreqLogit	5	0.25	1	0.25	2.736%	1	0.75	1	0.513%	0.75	0	0.5	0.522%
FreqRandom	1	0.75	0.75	0.5	2.753%	1	0	0	0.526%	0.25	1	0.5	0.525%
FreqRandom	2	1	0	0.5	2.741%	0.5	0.25	0	0.521%	0.75	0	0	0.525%
FreqRandom	3	0.75	0	0	2.739%	0.25	0.5	0.5	0.520%	1	0.75	0.25	0.524%
FreqRandom	4	0.5	1	0.75	2.736%	0.75	0.5	1	0.519%	0.5	0	1	0.522%
FreqRandom	5	0.75	1	0.75	2.736%	0	0.25	0	0.518%	1	0.5	0.5	0.520%

Table 7.3: Best average weekly F-measure, for each of the three graph topologies: all previous, creator and last post. For each of the four models, the five best combinations of parameters are shown, ordered by decreasing F-measure, expressed as a percentage.

		All previous				Creator				Last post			
Model	Rank	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F
FreqLCA	1	0.75	0.75	0.75	7.413%	0	0.5	0.75	1.764%	1	0.5	0.5	1.487%
FreqLCA	2	0.5	0	0.25	7.403%	0.5	0	0.25	1.751%	0.5	1	0.5	1.473%
FreqLCA	3	0	0.5	0	7.399%	0.75	0	0	1.751%	1	1	1	1.471%
FreqLCA	4	0.5	0	0	7.399%	0.5	0	0	1.747%	0	1	1	1.469%
FreqLCA	5	1	0.5	0.5	7.395%	0	0	0.25	1.741%	0.75	0.5	0.5	1.469%
FreqMax	1	0.25	0.5	0.25	12.434%	0.5	0.5	0.25	1.694%	0.5	0.5	0.25	2.065%
FreqMax	2	0.5	0.5	0.25	12.433%	0	0.25	1	1.691%	0.25	0.75	0.5	2.048%
FreqMax	3	0.25	0	1	12.426%	0	0.25	0.75	1.687%	1	0.75	0.5	2.048%
FreqMax	4	0.25	1	0.25	12.424%	1	1	0.5	1.687%	0.5	0.25	0.25	2.046%
FreqMax	5	0.75	1	0.25	12.424%	0.75	0.5	0.25	1.686%	0	1	0.75	2.045%
FreqLogit	1	1	1	1	7.485%	0.25	1	0.25	1.762%	0	0.5	0	1.520%
FreqLogit	2	0.25	1	0.25	7.463%	0.25	0.75	0.5	1.755%	0.75	0.25	0.75	1.501%
FreqLogit	3	0.75	0.25	0.75	7.458%	1	0.75	1	1.753%	1	1	0.5	1.496%
FreqLogit	4	0.75	0.75	1	7.457%	1	0	0	1.750%	0.25	0.5	0.25	1.495%
FreqLogit	5	0	0.25	0.25	7.446%	0.5	0.5	0.5	1.744%	1	0	0.25	1.490%
FreqRandom	1	0.75	1	0.75	7.473%	0.75	0.25	0.5	1.764%	1	0.25	1	1.506%
FreqRandom	2	0.75	0.5	0.75	7.468%	0.25	0.75	0	1.760%	0.75	1	0.75	1.506%
FreqRandom	3	1	0	0.5	7.467%	0.25	0	0.25	1.754%	0.5	0.5	0.25	1.504%
FreqRandom	4	0.5	0	0	7.456%	0.5	0.25	0	1.753%	0.25	0.5	0.5	1.502%
FreqRandom	5	0.25	0.75	0.5	7.454%	0.5	0.5	0.25	1.753%	0	0.75	0.5	1.498%

Table 7.4: Best average monthly F-measure, for each of the three graph topologies: all previous, creator and last post. For each of the four models, the five best combinations of parameters are shown, ordered by decreasing F-measure, expressed as a percentage.

the all previous topology (FreqLogit is the best with 20.997%), despite it consistently outperformed the other models for lower time scales. Still for the all previous topology, very little difference is observed between FreqLCA, FreqLogit and FreqRandom models. The same observations apply to the creator topology: FreqMax is outperformed (FreqRandom is the best with 7.956%), with very little variation between best models. For the last post topology, the FreqMax has the best performance ($F = 8.087\%$), with very little variation again between the other models' performance.

Greater regularities with respect to parameters values emerge. c_π values tend to be 0.5 or more across models for the all previous topology, and c_ν tends to be 0.5 or less across models for the last reply topology. By equation 5.4 a high value of c_π implies a small influence of users' preferences upon publications (users are more community oriented). By equation 5.5 a small value of c_ν means that social image is less affected by a publication. However, these results are not observed across topologies. In this sense, the most consistent model is FreqMax, since it tends to exhibit a c_μ parameter equal to 0.5 or less for all topologies.

		All previous				Creator				Last post			
Model	Rank	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F	c_π	c_μ	c_ν	F
FreqLCA	1	0.75	0.5	0.25	20.854%	0.5	0	1	7.911%	0.25	0.75	0	7.341%
FreqLCA	2	0.75	0.5	0.5	20.839%	1	0.5	0	7.898%	0.75	0.5	0.5	7.337%
FreqLCA	3	0.5	0	0	20.827%	0.75	0	0.75	7.881%	0	1	0.25	7.323%
FreqLCA	4	0.5	0.75	0.75	20.822%	0	0.5	0.25	7.879%	0.25	0.5	0.25	7.317%
FreqLCA	5	0.5	0.75	0.25	20.801%	0.75	0.25	0.75	7.877%	0.5	0.75	0.25	7.309%
FreqMax	1	1	0.25	0	19.921%	0.5	0.25	1	3.624%	0	0.25	0	8.087%
FreqMax	2	0	1	0	19.920%	0	0.25	1	3.621%	1	0.25	0	8.080%
FreqMax	3	0.75	0	0.5	19.918%	0	0.25	0.5	3.621%	0.5	0.25	0.5	8.076%
FreqMax	4	0	0	0	19.918%	0.25	0.25	0.75	3.619%	0.5	0.5	0.25	8.056%
FreqMax	5	0.75	0	0.75	19.917%	0	0.25	0.75	3.619%	0	0.5	0	8.044%
FreqLogit	1	0	0	0.25	20.997%	0.25	1	0.25	7.934%	0	0	0.25	7.418%
FreqLogit	2	0.75	1	0.5	20.993%	0	0	0.25	7.907%	0.75	0.25	0.75	7.415%
FreqLogit	3	0.75	1	0.25	20.985%	0.75	0.25	0.75	7.903%	0.25	1	0.25	7.413%
FreqLogit	4	0.75	0.5	1	20.972%	0.75	1	0.5	7.900%	0.75	0.5	1	7.412%
FreqLogit	5	0.75	0.25	0.75	20.961%	0.5	0.5	0.5	7.897%	0.5	1	0.5	7.403%
FreqRandom	1	0.75	0.5	0.25	20.990%	0	0.75	0.75	7.956%	0.75	0.5	0.25	7.447%
FreqRandom	2	0.75	0.5	0.75	20.985%	0.75	0.75	0.75	7.943%	0	0.5	0	7.434%
FreqRandom	3	0.5	0.25	0	20.982%	0.5	0.75	0.75	7.930%	0.25	0.75	0	7.430%
FreqRandom	4	0.25	0	0.5	20.961%	0.75	0.75	0	7.925%	1	0	0.5	7.426%
FreqRandom	5	0.75	0.75	0.75	20.957%	0.75	1	1	7.924%	1	0.5	0.5	7.422%

Table 7.5: Best average global F-measure, for each of the three graph topologies: all previous, creator and last post. For each of the four models, the five best combinations of parameters are shown, ordered by decreasing F-measure, expressed as a percentage.

7.3 Contents variance

Information diffusion is measured through the temporal trend of contents variance among users (section 6.4.3). If a convergence of users towards a common opinion occurs, therefore

Model	c_π	c_μ	c_ν	$\hat{b}$
FreqLCA	0	1	0.5	7.32E-007
FreqLCA	0	1	0.75	4.34E-007
FreqLCA	0.5	0.25	0	-1.69E-006
FreqLCA	0.5	0.75	0	1.52E-006
FreqLCA	0.5	0.5	1	-1.16E-006
FreqMax	0.75	0	0.25	1.09E-005
FreqMax	0	0	0.75	2.67E-006
FreqMax	0.25	0	1	3.60E-006
FreqMax	0.75	0	0	1.46E-005
FreqMax	0.75	0	1	1.36E-005
FreqLogit	1	0.75	0.5	4.45E-006
FreqLogit	0.25	0.5	0.75	-8.35E-007
FreqLogit	0.25	0.5	0.5	-1.54E-006
FreqLogit	1	0.25	0.75	2.81E-006
FreqLogit	0.5	0.75	0.25	2.16E-006
FreqRandom	1	0.75	0.5	1.78E-006
FreqRandom	1	0.5	0	4.08E-006
FreqRandom	1	0.5	0.25	2.88E-006
FreqRandom	0.5	1	0.5	2.99E-006
FreqRandom	0.25	0.5	0.75	-1.23E-006

Table 7.6: Temporal trends of mean topic weights variance for the best models discussed in section 7.1.1 (table 7.1).

the variance between the topic weights of published posts should converge to zero. To detect this convergence, the OLS estimator of the weekly mean standard deviation over topics and users is computed. Positive trends in the mean topic weights variance are observed in the results though, rather than negative (table 7.6). Therefore, during the period of simulation no convergence to a common opinion is observed. Complete results are found in appendix E.

7.4 Discussion

Information gains around 0.4% are obtained at the weekly level, and around 0.6% at the monthly level. Therefore, the models improve the MAPE obtained by a flat pattern of publication. However, the model with the greatest information gain is the random model, which does not consider any information for the choice of posts and threads. This means that the information considered by the models FreqLCA, FreqLogit and FreqMax does not improve adjustment. Surprisingly, the FreqMax, which theoretically optimizes its decisions, yields the worse results. Nonetheless, it is also the model with globally the best results in graph generation, though with very low levels of F-measure. One explanation is that the right contents are published in the wrong thread, thus provoking bad interaction graphs. The simplified structure of the modeled forum may have some responsibility in this fact. Indeed, in the case of the FreqMax model, the thread that maximizes overall utility is chosen, no matter of how old it is. Therefore, the simulation make users may post in places in which they would not in real conditions.

An important drawback is the lack of client-side analysis of the results by the administrators of Plexilandia. Indeed, it has not been possible to obtain the results presented above in time for discussion. The above are therefore technical results that need user validation, as stated in the initial objectives of this work.

Chapter 8

Conclusions

This work is an attempt to implement a realistic information diffusion model through an online social network. There exists an extensive literature on diffusion research in general terms, and innovations diffusion research provides a particularly appropriated framework. Indeed, innovations diffusion is combining with ever greater success the mathematics of diffusion with the social nature of the diffusion agents involved. Further, there is an incredible amount of measured evidence since the advent of the Internet. In their quality of social beings themselves, humans are privileged observers of social phenomena, including social diffusion processes. This is an important advantage, as physical diffusion processes, for example, must be measured by much more indirect means. The social observer is himself the instrument of measure: his own cognitive functions have evolved to survive in a world that is both physical and social. However, this advantage may also be a weakness: objectivity is more difficult to achieve, and the link between mathematics and social intuition is a challenge on its own.

Social diffusion is mathematically modeled since forty years: it is a young field of research. It lies mainly in the intersection of physics, mathematics, epidemiology, social science and psychology: it is interdisciplinary by nature. Further, the inclusion of network structures in the research scope has brought an increasing complexity in the study of statistical properties in diffusion processes. Therefore, it has been deliberately decided to perform a review of basic theoretical background, the very foundations of diffusion processes from a social sciences perspective. Early work in social sciences, early mathematical models of diffusion, decision making models and information modeling have been discussed, in order to build a comprehensive framework of analysis. It is hoped that this approach may enable more advanced future work.

The social diffusion process studied in this work consists of web forum activity. A general forum-agent framework has been defined, and a web forum simulator has been implemented. In the context of the general framework, four specific models of user behavior have been defined and implemented. One of them — the main model — is based on a novel model of perceptual choice. The three other models include variants of the voter model. Web forum activity has been simulated with these four models and compared with real data, across three

dimensions of analysis: contents generation, graph generation, and contents variance. The main model has been found to be consistently outperformed by two models, and therefore the first hypothesis of this work has not been validated. Moreover, no detection of opinion convergence has been detected, which might have been caused by both too short a simulation time window and poor metric choice. Hence, neither the second hypothesis of this work has been validated. All initial objectives have been met, except for the practical usefulness assessment, due to time constraints.

However, the implemented classes and the methodological framework provide a sound basis for future work, and room is available for results improvement. The most important modification is a more accurate modeling of the web forum structure, since it is thought to be a major cause in the models' poor results. A broader exploration of the models parameters, in the particular for the LCA-based model, might be instructive too. Finally, an interesting approach is to separate the diffusion process into topic-specific processes, as each topic may have its own dynamics. As a general recommendation for future work, it seems mandatory to perform a more technical study, with better diffusion metrics.

From a personal perspective, this work means the end of an important phase, and the beginning of another. It is the result of one year of work, and it is hoped that the more relevant findings are successfully written down. An important challenge has been to maintain consistency, in conceptual and working terms. To achieve this, it is very important to perform a good literature review from the very start, while progressively increasing practical work. A good theoretical framework is crucial when things go wrong and the results are not as good as expected. Moreover, it provides a language that enables common understanding for the diffusion research community. Practical effort has higher chances to achieve good results if the fundamental concepts are well understood.

Research should not be considered as separated from the entrepreneurial world. On the contrary, research should emerge as a consequence of technical and operational excellence: once practical problems are mastered, there is room for thinking and inventing new solutions for new business problems. In this sense, it has been an extremely exciting experience to perform a research study, while keeping in touch with online marketing practitioners. The author of this work remains very grateful.

Appendix

Appendix A

Notes on Bass' paper

In the original paper *A New Product Growth Model For Consumer Durables* ([16]), a calculation was not explained, and some errors were found. In this document, the unexplained calculation will be see in detail, and the errors will be corrected.

A.1 Calculation of the expected time to purchase

Let T be the continuous random variable "time to purchase", then its expectation is by definition:

$$E[T] = \int_0^{+\infty} sf(s)\mathrm{d}s = \lim_{t\to+\infty} \int_0^t sf(s)\mathrm{d}s \tag{A.1}$$

where $f(.)$ is the density function of T. Integrating by parts:

$$E[T] = \lim_{t\to+\infty} \left(\left[sF(s) \right]_0^t - \int_0^t F(s)\mathrm{d}s \right) \tag{A.2}$$

where $F(.)$ is the distribution function of T ($\frac{\mathrm{d}F(t)}{\mathrm{d}t} = f(t)$). From [16] (p. 218) it is known that:

$$F(t) = \frac{1 - \mathrm{e}^{-(p+q)t}}{\frac{q}{p}\mathrm{e}^{-(p+q)t} + 1} \tag{A.3}$$

Rewriting (A.3):

$$F(t) = \frac{1 + \frac{q}{p}\mathrm{e}^{-(p+q)t} - \frac{q}{p}\mathrm{e}^{-(p+q)t} - \mathrm{e}^{-(p+q)t}}{\frac{q}{p}\mathrm{e}^{-(p+q)t} + 1} \tag{A.4}$$

$$= 1 - (\frac{q}{p} + 1)\frac{\mathrm{e}^{-(p+q)t}}{\frac{q}{p}\mathrm{e}^{-(p+q)t} + 1} \tag{A.5}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t}\left(t + \frac{1}{q}\ln\left(\frac{q}{p}\mathrm{e}^{-(p+q)t} + 1\right)\right) \tag{A.6}$$

Replacing (A.5) and (A.6) in (A.2):

$$E[T] = \lim_{t\to+\infty}\left(\left[s(1-(\frac{q}{p}+1)\frac{\mathrm{e}^{-(p+q)s}}{\frac{q}{p}\mathrm{e}^{-(p+q)s}+1})\right]_0^t - \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\left(s+\frac{1}{q}\ln\left(\frac{q}{p}\mathrm{e}^{-(p+q)s}+1\right)\right)\mathrm{d}s\right) \quad \text{(A.7)}$$

$$= \lim_{t\to\infty}\left(t-(\frac{q}{p}+1)\frac{t}{\frac{q}{p}+\mathrm{e}^{-(p+q)t}} - \left[s+\frac{1}{q}\ln\left(\frac{q}{p}\mathrm{e}^{-(p+q)s}+1\right]_0^t\right) \quad \text{(A.8)}$$

$$= \lim_{t\to\infty}\left(t-(\frac{q}{p}+1)\frac{t}{\frac{q}{p}+\mathrm{e}^{-(p+q)t}} - t - \frac{1}{q}\ln\left(\frac{q}{p}\mathrm{e}^{-(p+q)t}+1\right) + \frac{1}{q}\ln\left(\frac{q}{p}+1\right)\right) \quad \text{(A.9)}$$

$$= \frac{1}{q}\ln\left(\frac{q}{p}+1\right) \quad \text{(A.10)}$$

$$= \frac{1}{q}\ln\left(\frac{p+q}{p}\right) \quad \text{(A.11)}$$

as $\lim_{t\to\infty}\frac{t}{\frac{q}{p}+\mathrm{e}^{-(p+q)t}} = 0$ and $\lim_{t\to\infty}\ln\left(\frac{q}{p}\mathrm{e}^{-(p+q)t}+1\right) = 0$. The result in eq. (A.11) is the same as found in [16], page 213.

A.2 Errors in page 223

The first error is $\sum_{t=0}^{x-1} F(x)/f(t) = k$, since the correct expression is $F(x)/\sum_{t=0}^{x-1} f(t) = k$. Next, the exponential distribution with event rate λ has a density function equal to $f(x) = \lambda\mathrm{e}^{-\lambda x}$ and a distribution function equal to $F(x) = 1-\mathrm{e}^{-\lambda x}$. Note that:

$$F(x+1) - F(x) = 1 - \mathrm{e}^{-\lambda(x+1)} - (1-\mathrm{e}^{-\lambda x}) \quad \text{(A.12)}$$

$$= \mathrm{e}^{-\lambda x} - \mathrm{e}^{-\lambda x}\mathrm{e}^{-\lambda} \quad \text{(A.13)}$$

$$= \mathrm{e}^{-\lambda x}(1-\mathrm{e}^{-\lambda}) \quad \text{(A.14)}$$

Therefore the density function may be rewritten as:

$$f(x) = \frac{\lambda}{1-\mathrm{e}^{-\lambda}}\big(F(x+1) - F(x)\big) \quad \text{(A.15)}$$

Supposing that when p and T are small the density function of T is approximately exponential with rate $p+q$, then $1/k$ should be equal to $(p+q)/(1-\mathrm{e}^{-(p+q)})$, not $(p+q)/(\mathrm{e}^{p+q}+1)$. Finally, the correct expression for q is:

$$q = \frac{0.97q'}{1+0.4(1+1/\theta)q'} \quad \text{(A.16)}$$

Appendix B

Deduction of logit probabilities

According to Guadagni and Little ([26]), the assumptions of the logit model are:

"(1) Alternative $k \in S_\mathrm{i}$ holds for the individual a preference or *utility*,

$$u_k = v_k + \varepsilon_k \quad , \qquad \text{where} \tag{B.1}$$

v_k = a deterministic component of i's utility, to be calculated from observed variables, and ε_k = a random component of i's utility, varying from choice occasion to choice occasion, possibly as a result of unobserved variables.
(2) Confronted by the set of alternatives, individual i chooses the one with the highest utility on the occasion. I.e., the probability of choosing k is

$$p_k = P\{u_k \geq u_j, j \in S_\mathrm{i}\} \tag{B.2}$$

(3) The $\varepsilon_k, k \in S_\mathrm{i}$, are independently distributed random variables with a double exponential (Gumbel type II extreme value) distribution

$$P(\varepsilon_k \leq \varepsilon) = \mathrm{e}^{-\mathrm{e}^{-\varepsilon}} \quad , \qquad -\infty < \varepsilon < \infty \quad \text{"([26], p. 207)} \tag{B.3}$$

The choice probability for alternative $k \in S_\mathrm{i}$ will be deduced from equations B.1, B.2 and B.3.

Equation B.3 defines the distribution function $F(\varepsilon) = \mathrm{e}^{-\mathrm{e}^{-\varepsilon}}$, and therefore the density function of ε_k is:

$$f(\varepsilon) = \frac{\mathrm{d}F(\varepsilon)}{\mathrm{d}\varepsilon} = \mathrm{e}^{-\mathrm{e}^{-\varepsilon}}\mathrm{e}^{-\varepsilon} \quad \text{with} \quad \int_{-\infty}^{\infty} f(\varepsilon)\mathrm{d}\varepsilon = F(\varepsilon)\big|_{-\infty}^{\infty} = 1 \tag{B.4}$$

From equation B.2:

$$\begin{aligned} p_k &= P\{u_k \geq u_j, j \in S_\mathrm{i}\} \\ &= P\{v_k + \varepsilon_k \geq v_j + \varepsilon_j, j \in S_\mathrm{i}\} \quad \text{by B.1} \\ &= P\{\varepsilon_j \leq v_k - v_j + \varepsilon_k, j \in S_\mathrm{i}\} \end{aligned} \tag{B.5}$$

Integrating over ε_k, equation B.5 becomes:

$$\begin{aligned}
p_k &= \int_{-\infty}^{\infty} P[\varepsilon_j \leq v_k - v_j + \varepsilon_k, j \in S_\mathrm{i}] f(\varepsilon_k) \mathrm{d}\varepsilon_k \\
&= \int_{-\infty}^{\infty} P[\varepsilon_j \leq v_k - v_j + \varepsilon_k, j \in S_\mathrm{i}] \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k}} \mathrm{e}^{-\varepsilon_k} \mathrm{d}\varepsilon_k \quad \text{by B.4} \\
&= \int_{-\infty}^{\infty} \prod_{j \neq k} \mathrm{e}^{-\mathrm{e}^{-(v_k - v_j + \varepsilon_k)}} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k}} \mathrm{e}^{-\varepsilon_k} \mathrm{d}\varepsilon_k \\
&= \int_{-\infty}^{\infty} \mathrm{e}^{-\sum_{j \neq k} \mathrm{e}^{-(v_k - v_j + \varepsilon_k)}} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k}} \mathrm{e}^{-\varepsilon_k} \mathrm{d}\varepsilon_k \\
&= \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k} \sum_{j \neq k} \mathrm{e}^{v_j - v_k}} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k}} \mathrm{e}^{-\varepsilon_k} \mathrm{d}\varepsilon_k \\
&= \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{e}^{-\varepsilon_k}(1 + \sum_{j \neq k} \mathrm{e}^{v_j - v_k})} \mathrm{e}^{-\varepsilon_k} \mathrm{d}\varepsilon_k
\end{aligned} \tag{B.6}$$

Let $\alpha = 1 + \sum_{j \neq k} \mathrm{e}^{v_j - v_k}$, then the change of variable $\alpha \mathrm{e}^{-\varepsilon_k} = \mathrm{e}^{-s}$ yields $\mathrm{d}\varepsilon_k = \mathrm{d}s$, and equation B.6 becomes:

$$\begin{aligned}
p_k &= \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{e}^{-s}} \frac{\mathrm{e}^{-s}}{\alpha} \mathrm{d}s \\
&= \frac{1}{\alpha} \int_{-\infty}^{\infty} f(s) \mathrm{d}s \quad \text{by B.4} \\
&= \frac{1}{1 + \sum_{j \neq k} \mathrm{e}^{v_j - v_k}} \quad \text{by B.4}
\end{aligned} \tag{B.7}$$

Finally, multiplying the numerator and the denominator by e^{v_k}, equation B.7 becomes:

$$p_k = \frac{\mathrm{e}^{v_k}}{\sum_{j \in S_\mathrm{i}} \mathrm{e}^{v_j}} \tag{B.8}$$

Appendix C

Contents generation results

	Weekly	Monthly
FreqLCA	page 110	page 114
FreqMax	page 111	page 115
FreqLogit	page 112	page 116
FreqRandom	page 113	page 117

Table C.1: MAPE results index.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0	1	0.5	7.321%	46	0.75	1	0	7.346%	91	0.5	1	1	7.360%
2	0	1	0.75	7.323%	47	1	0.75	0.25	7.346%	92	0.5	0.25	0.5	7.361%
3	0.5	0.25	0	7.326%	48	0.75	0.25	1	7.346%	93	1	0	0.25	7.361%
4	0.5	0.75	0	7.327%	49	0	0.25	0.25	7.346%	94	0	0.5	0.25	7.361%
5	0.5	0.5	1	7.329%	50	0.75	1	1	7.346%	95	1	0	0.5	7.361%
6	0.25	1	0.25	7.331%	51	0	0.25	0	7.346%	96	0.25	0.25	0	7.361%
7	0	0.75	0	7.332%	52	0	0.75	0.5	7.346%	97	0.25	1	1	7.362%
8	0	1	0.25	7.332%	53	1	1	0	7.346%	98	0.5	0.75	1	7.363%
9	0.75	0.75	0	7.333%	54	1	1	1	7.347%	99	1	0.25	0	7.364%
10	0.5	0.75	0.5	7.334%	55	0.5	0.5	0	7.347%	100	1	0	0.75	7.365%
11	0.25	0.5	0.25	7.335%	56	1	0.25	0.25	7.347%	101	1	0.5	0	7.366%
12	0.75	0.25	0.25	7.335%	57	0.75	0.5	0.25	7.348%	102	0.75	1	0.25	7.366%
13	0.75	0.75	0.5	7.337%	58	0.25	0.75	0	7.348%	103	0.75	0.25	0.75	7.366%
14	0.75	1	0.5	7.337%	59	0.25	1	0.75	7.349%	104	1	0	0	7.367%
15	0.75	0	0.25	7.337%	60	0.25	0.75	1	7.349%	105	0.5	0.25	1	7.368%
16	0	1	1	7.337%	61	0.75	0.75	1	7.350%	106	0.5	0.5	0.25	7.369%
17	0.25	0.75	0.75	7.338%	62	0.25	0.5	0.75	7.350%	107	1	0	1	7.371%
18	0.25	0.25	0.25	7.338%	63	0.75	0	0.5	7.352%	108	1	0.5	0.25	7.372%
19	1	1	0.75	7.338%	64	0.75	0.5	0.5	7.352%	109	1	0.5	0.5	7.378%
20	0.75	0	1	7.339%	65	0.5	0.25	0.75	7.352%	110	0.25	1	0	7.378%
21	1	0.75	0.75	7.339%	66	0.5	0.5	0.5	7.353%	111	0.5	0	0.25	7.394%
22	0.5	1	0.5	7.339%	67	1	1	0.25	7.353%	112	0.5	0	0	7.396%
23	0.25	0.75	0.25	7.339%	68	0	0.25	1	7.353%	113	0.5	0	0.75	7.409%
24	0.75	0.75	0.25	7.339%	69	0.75	0.5	0.75	7.353%	114	0.5	0	1	7.412%
25	0	0.75	0.25	7.340%	70	0	0.75	1	7.353%	115	0.5	0	0.5	7.416%
26	0.25	0.5	0	7.340%	71	0.75	0.25	0	7.354%	116	0.25	0	0.5	7.541%
27	0	0.25	0.5	7.341%	72	0.75	0.75	0.75	7.354%	117	0.25	0	0.75	7.542%
28	0.5	1	0	7.341%	73	0.5	0.5	0.75	7.354%	118	0.25	0	1	7.543%
29	0	0.5	1	7.342%	74	0.5	1	0.25	7.355%	119	0.25	0	0.25	7.543%
30	1	1	0.5	7.342%	75	1	0.25	1	7.356%	120	0.25	0	0	7.551%
31	0.75	0	0.75	7.342%	76	0.75	0	0	7.356%	121	0	0	0.5	7.710%
32	0	1	0	7.343%	77	1	0.75	0.5	7.357%	122	0	0	1	7.713%
33	1	0.75	1	7.343%	78	0.5	0.25	0.25	7.357%	123	0	0	0.75	7.722%
34	1	0.5	1	7.343%	79	0.25	1	0.5	7.357%	124	0	0	0	7.729%
35	0.25	0.5	1	7.343%	80	0.75	1	0.75	7.357%	125	0	0	0.25	7.736%
36	0.75	0.5	1	7.343%	81	1	0.25	0.5	7.357%					
37	1	0.5	0.75	7.344%	82	0.25	0.25	1	7.357%					
38	0.25	0.5	0.5	7.344%	83	1	0.25	0.75	7.357%					
39	0	0.75	0.75	7.344%	84	0.75	0.25	0.5	7.357%					
40	0	0.5	0.75	7.344%	85	1	0.75	0	7.358%					
41	0	0.5	0.5	7.345%	86	0.75	0.5	0	7.358%					
42	0	0.25	0.75	7.345%	87	0.25	0.75	0.5	7.359%					
43	0.25	0.25	0.75	7.345%	88	0.5	1	0.75	7.359%					
44	0	0.5	0	7.346%	89	0.5	0.75	0.75	7.359%					
45	0.5	0.75	0.25	7.346%	90	0.25	0.25	0.5	7.359%					

Table C.2: Average weekly MAPE, model FreqLCA.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0.75	0	0.25	7.693%	46	0	1	0.25	7.817%	91	0	0.75	0.75	7.840%
2	0	0	0.75	7.693%	47	0.75	0.5	0.75	7.818%	92	0.25	1	1	7.841%
3	0.25	0	1	7.695%	48	0.5	0.5	0.75	7.818%	93	0	1	0.75	7.841%
4	0.75	0	0	7.709%	49	0.25	0.5	1	7.819%	94	1	0.5	0.25	7.841%
5	0.75	0	1	7.710%	50	0.75	0.25	0.75	7.821%	95	1	1	0.25	7.841%
6	0	0	0	7.710%	51	0	0.5	0.25	7.821%	96	1	0.25	0.5	7.841%
7	0	0	0.5	7.711%	52	0.5	0.75	0.25	7.822%	97	0.5	1	0.5	7.842%
8	0.25	0	0.75	7.712%	53	0	0.5	0	7.823%	98	1	0	0.25	7.843%
9	0	0	0.25	7.714%	54	0	0.5	0.5	7.824%	99	1	0.5	0.5	7.845%
10	0.5	0	0.5	7.715%	55	0.25	0.5	0.5	7.824%	100	0.5	1	0.75	7.845%
11	0.25	0	0.5	7.717%	56	1	0.25	0.75	7.825%	101	0.25	1	0.25	7.845%
12	0.25	0	0	7.721%	57	1	0.25	0.25	7.826%	102	1	0.5	0.75	7.846%
13	0.5	0	0.25	7.724%	58	0	1	0.5	7.826%	103	0.75	0.5	0.25	7.846%
14	0.75	0	0.5	7.725%	59	0.75	0.75	0.75	7.826%	104	1	0.75	0.75	7.846%
15	0.5	0	0	7.725%	60	1	0	0.75	7.826%	105	0.5	0.75	0	7.847%
16	0.75	0	0.75	7.726%	61	0.25	0.75	0.75	7.826%	106	0.75	0.75	1	7.847%
17	0	0	1	7.728%	62	0.75	0.25	0.25	7.826%	107	0.75	0.75	0	7.848%
18	0.25	0	0.25	7.738%	63	0.75	0.5	0.5	7.827%	108	0.25	0.75	0	7.849%
19	0.5	0	0.75	7.741%	64	0	0.5	0.75	7.828%	109	1	1	0	7.849%
20	0.5	0	1	7.757%	65	0	0.75	0.5	7.828%	110	0.5	0.75	0.5	7.849%
21	0.25	0.25	1	7.767%	66	0.25	0.5	0.75	7.828%	111	0.75	0.5	1	7.850%
22	0.25	0.25	0.25	7.776%	67	0.25	0.5	0.25	7.828%	112	0.75	1	0.75	7.851%
23	0	0.25	0	7.779%	68	0.5	0.75	1	7.829%	113	1	0.75	1	7.851%
24	0.25	0.25	0.75	7.783%	69	0.5	0.75	0.75	7.829%	114	0.25	1	0.5	7.852%
25	0	0.25	0.5	7.784%	70	0	0.75	0	7.830%	115	0.75	0.75	0.5	7.853%
26	0.25	0.25	0.5	7.784%	71	0.75	1	0.5	7.830%	116	1	0.25	0	7.853%
27	0	0.25	1	7.787%	72	1	1	1	7.831%	117	1	0.25	1	7.856%
28	0.25	0.25	0	7.791%	73	1	0.5	1	7.832%	118	1	0.75	0.5	7.858%
29	0	0.25	0.25	7.794%	74	0.5	0.5	0.25	7.832%	119	0.5	1	0	7.861%
30	0	0.25	0.75	7.795%	75	0.75	1	0.25	7.832%	120	1	1	0.5	7.861%
31	0.5	0.25	1	7.797%	76	1	0	0.5	7.833%	121	1	0.5	0	7.862%
32	0.5	0.5	1	7.802%	77	1	0.75	0	7.833%	122	0.75	1	0	7.862%
33	0.5	0.25	0.25	7.803%	78	0.25	0.75	0.25	7.833%	123	0.25	1	0	7.865%
34	0.5	0.25	0.75	7.804%	79	0.75	1	1	7.834%	124	1	0	0	7.865%
35	0.75	0.25	1	7.805%	80	1	1	0.75	7.834%	125	0	1	0	7.871%
36	0.5	0.25	0	7.806%	81	0.5	0.5	0	7.835%					
37	0	0.5	1	7.807%	82	0	0.75	0.25	7.835%					
38	0.5	0.5	0.5	7.807%	83	0.75	0.5	0	7.835%					
39	0.75	0.25	0.5	7.807%	84	0.25	0.75	0.5	7.836%					
40	0	0.75	1	7.808%	85	0.5	1	0.25	7.837%					
41	0.75	0.25	0	7.808%	86	1	0	1	7.838%					
42	0.5	0.25	0.5	7.810%	87	0.5	1	1	7.838%					
43	0.25	0.5	0	7.811%	88	0.75	0.75	0.25	7.839%					
44	0.25	1	0.75	7.812%	89	1	0.75	0.25	7.839%					
45	0.25	0.75	1	7.813%	90	0	1	1	7.839%					

Table C.3: Average weekly MAPE, model FreqMax.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	1	0.75	0.5	7.272%	46	0	0.5	0.5	7.297%	91	0.75	0.5	0.25	7.310%
2	0.25	0.5	0.75	7.275%	47	0.75	0.75	0.5	7.297%	92	0.75	0.75	0.75	7.311%
3	0.25	0.5	0.5	7.276%	48	0.75	0	1	7.298%	93	0	1	0.75	7.311%
4	1	0.25	0.75	7.281%	49	0.5	1	0.75	7.298%	94	0.75	0	0.5	7.311%
5	0.5	0.75	0.25	7.281%	50	0.75	0.5	1	7.299%	95	0.5	0.5	1	7.311%
6	0.25	0.75	0	7.282%	51	0.25	0.75	0.75	7.299%	96	0.25	0.75	1	7.312%
7	0	1	0.5	7.282%	52	0.25	0.5	0	7.299%	97	1	0.75	1	7.312%
8	1	0.5	0.25	7.283%	53	0.75	0.25	1	7.299%	98	1	0.5	0.75	7.313%
9	0	0.25	0.5	7.283%	54	0.25	1	0.25	7.299%	99	1	0.5	0.5	7.313%
10	0	0.5	0.75	7.284%	55	0.25	1	0.5	7.299%	100	0	1	0.25	7.314%
11	0.75	0.25	0.25	7.284%	56	0.25	0.25	0.25	7.299%	101	1	0	0.75	7.314%
12	0	0.75	0.25	7.284%	57	0.75	0.75	1	7.299%	102	0	0.75	0	7.316%
13	0.75	1	0	7.285%	58	0.75	0.25	0.75	7.299%	103	0	0.5	0	7.317%
14	0.5	1	1	7.286%	59	0.5	0.25	0	7.300%	104	0.25	1	0	7.317%
15	0	0.75	0.75	7.287%	60	0.5	0.75	0.5	7.300%	105	0.75	0.5	0.75	7.317%
16	0.5	0.75	0	7.287%	61	0.5	0.25	0.5	7.301%	106	1	1	1	7.318%
17	0.75	0.5	0.5	7.288%	62	1	0	0.25	7.301%	107	0.75	1	0.25	7.321%
18	0.75	0	0.75	7.288%	63	0.25	1	1	7.301%	108	1	0	0.5	7.323%
19	0.5	0.25	1	7.288%	64	0.5	1	0.5	7.301%	109	0	1	1	7.323%
20	0	0.25	0.25	7.288%	65	1	1	0	7.301%	110	1	1	0.5	7.324%
21	0.25	0.75	0.25	7.289%	66	0.25	0.25	0	7.301%	111	0.5	0	1	7.375%
22	0.75	0.25	0.5	7.289%	67	0.75	0.5	0	7.301%	112	0.5	0	0.25	7.387%
23	0	1	0	7.289%	68	0.75	0	0.25	7.302%	113	0.5	0	0.5	7.392%
24	0.75	0.25	0	7.289%	69	0.25	0.5	1	7.302%	114	0.5	0	0	7.393%
25	0.25	0.25	1	7.289%	70	0	0.25	0	7.302%	115	0.5	0	0.75	7.397%
26	1	0.5	0	7.289%	71	0	0.25	0.75	7.302%	116	0.25	0	0.5	7.534%
27	0.5	0.25	0.25	7.290%	72	1	0.75	0.75	7.302%	117	0.25	0	0	7.542%
28	1	0.5	1	7.291%	73	1	1	0.25	7.303%	118	0.25	0	0.25	7.547%
29	0	0.75	1	7.292%	74	0.25	0.75	0.5	7.304%	119	0.25	0	1	7.550%
30	0.75	1	1	7.292%	75	0	0.5	1	7.304%	120	0.25	0	0.75	7.551%
31	1	0.25	1	7.292%	76	0.5	0.5	0.75	7.304%	121	0	0	0	7.702%
32	0.5	0.25	0.75	7.292%	77	0.75	0.75	0	7.304%	122	0	0	0.75	7.723%
33	0	0.25	1	7.292%	78	0.5	0.5	0.25	7.305%	123	0	0	1	7.726%
34	0.25	0.25	0.75	7.292%	79	0.5	0.5	0	7.305%	124	0	0	0.5	7.729%
35	0.5	1	0.25	7.292%	80	0.75	1	0.75	7.305%	125	0	0	0.25	7.732%
36	1	0.75	0.25	7.292%	81	0.5	0.5	0.5	7.305%					
37	0	0.5	0.25	7.292%	82	0.75	0.75	0.25	7.306%					
38	0.75	1	0.5	7.293%	83	0.5	0.75	0.75	7.306%					
39	0	0.75	0.5	7.293%	84	1	0.25	0	7.306%					
40	0.25	0.25	0.5	7.293%	85	0.5	1	0	7.307%					
41	0.5	0.75	1	7.293%	86	0.75	0	0	7.308%					
42	1	0	1	7.294%	87	1	0.75	0	7.309%					
43	0.25	1	0.75	7.295%	88	0.25	0.5	0.25	7.309%					
44	1	0.25	0.25	7.296%	89	1	1	0.75	7.310%					
45	1	0.25	0.5	7.296%	90	1	0	0	7.310%					

Table C.4: Average weekly MAPE, model FreqLogit.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	1	0.75	0.5	7.263%	46	0.25	1	0.5	7.292%	91	0.75	0	0.75	7.309%
2	1	0.5	0	7.274%	47	0.25	0.75	0.5	7.292%	92	1	0.25	0.5	7.310%
3	1	0.5	0.25	7.276%	48	0	0.25	1	7.293%	93	0.75	0.25	0.75	7.310%
4	0.5	1	0.5	7.276%	49	0.25	0.25	1	7.293%	94	0.75	1	0.25	7.310%
5	0.25	0.5	0.75	7.277%	50	0.75	0	0.25	7.293%	95	0.75	0.25	0.5	7.311%
6	0	0.25	0.25	7.277%	51	1	0.75	0.25	7.293%	96	0.5	0.75	1	7.311%
7	0	0.5	0	7.277%	52	0.75	0.5	0.25	7.294%	97	0.25	0.25	0.5	7.311%
8	1	1	0	7.278%	53	0.25	0.75	1	7.294%	98	1	0.75	0	7.311%
9	1	0.5	1	7.278%	54	0	0.5	0.25	7.295%	99	1	0.5	0.5	7.311%
10	1	1	1	7.278%	55	0	1	0	7.295%	100	1	0	0.25	7.312%
11	0.75	0.75	1	7.279%	56	1	1	0.75	7.295%	101	0.25	0.5	0.5	7.313%
12	0	1	0.75	7.279%	57	0.5	0.25	0.5	7.295%	102	0.75	0.75	0.75	7.314%
13	0.25	0.75	0.25	7.280%	58	0.25	0.5	1	7.295%	103	1	0.25	0.75	7.315%
14	1	0	0	7.280%	59	0.5	0.75	0.75	7.295%	104	0.75	0	1	7.316%
15	0.25	0.5	0.25	7.280%	60	0.75	0.5	0.75	7.295%	105	0.5	0.5	0.5	7.317%
16	0.75	1	1	7.280%	61	0.75	0.25	0.25	7.296%	106	0	0.75	0	7.318%
17	0.5	0.5	0	7.281%	62	1	0.25	0.25	7.296%	107	0	0.75	0.75	7.319%
18	0.5	1	0.25	7.281%	63	0.5	0.75	0	7.296%	108	0.75	0.75	0.25	7.322%
19	0	0.75	0.25	7.282%	64	0.25	0.25	0	7.296%	109	0.75	0	0.5	7.323%
20	0.25	0.75	0	7.282%	65	0.5	0.25	1	7.298%	110	0.5	0.5	1	7.325%
21	0.5	0.75	0.25	7.282%	66	0.25	0.75	0.75	7.298%	111	0.5	0	0.25	7.388%
22	0	0.25	0.5	7.283%	67	0	0.25	0.75	7.298%	112	0.5	0	1	7.408%
23	1	0.75	1	7.284%	68	0.25	0.25	0.25	7.298%	113	0.5	0	0	7.410%
24	0.75	0.5	0.5	7.285%	69	1	0.75	0.75	7.299%	114	0.5	0	0.5	7.411%
25	0.25	0.5	0	7.285%	70	0.5	0.5	0.75	7.299%	115	0.5	0	0.75	7.413%
26	1	0.25	1	7.285%	71	1	0.5	0.75	7.300%	116	0.25	0	0.5	7.544%
27	0	0.5	1	7.285%	72	0.75	1	0.5	7.300%	117	0.25	0	0.75	7.548%
28	0.75	0.5	0	7.286%	73	0	1	0.5	7.300%	118	0.25	0	0.25	7.551%
29	0.5	0.25	0	7.286%	74	0.75	1	0	7.300%	119	0.25	0	0	7.555%
30	1	1	0.5	7.286%	75	0.75	0.5	1	7.301%	120	0.25	0	1	7.564%
31	1	1	0.25	7.286%	76	0	0.5	0.5	7.301%	121	0	0	0.25	7.713%
32	0	0.5	0.75	7.287%	77	0	1	0.25	7.301%	122	0	0	0.75	7.717%
33	0	0.25	0	7.287%	78	0.75	0.75	0	7.302%	123	0	0	0.5	7.721%
34	0	1	1	7.287%	79	0.75	0	0	7.302%	124	0	0	0	7.722%
35	0.5	0.25	0.25	7.287%	80	0.5	0.5	0.25	7.302%	125	0	0	1	7.728%
36	1	0	0.5	7.287%	81	0.5	0.75	0.5	7.302%					
37	0.25	1	1	7.288%	82	1	0.25	0	7.303%					
38	0.75	0.25	0	7.288%	83	0.25	1	0.75	7.304%					
39	0.75	0.75	0.5	7.290%	84	0	0.75	1	7.306%					
40	1	0	0.75	7.290%	85	1	0	1	7.306%					
41	0.25	1	0	7.290%	86	0.75	1	0.75	7.306%					
42	0.25	0.25	0.75	7.291%	87	0.75	0.25	1	7.306%					
43	0.25	1	0.25	7.291%	88	0.5	1	0	7.306%					
44	0	0.75	0.5	7.291%	89	0.5	1	1	7.307%					
45	0.5	1	0.75	7.292%	90	0.5	0.25	0.75	7.308%					

Table C.5: Average weekly MAPE, model FreqRandom.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0.75	0	1	4.463%	46	0.5	0.25	0	4.513%	91	0.5	0.5	0	4.525%
2	0.75	0	0.25	4.470%	47	0.75	1	0.25	4.513%	92	1	0.5	1	4.525%
3	0.75	0	0	4.473%	48	0	0.75	0.25	4.513%	93	0.75	0.25	1	4.526%
4	0.75	0	0.5	4.474%	49	0.25	0.25	0.25	4.514%	94	0.75	0.25	0.5	4.527%
5	0.75	0	0.75	4.474%	50	0.5	1	0	4.514%	95	1	0.75	0.25	4.527%
6	0.5	0	0.25	4.485%	51	0.5	1	1	4.514%	96	0.5	0.25	1	4.527%
7	0.75	1	1	4.488%	52	0	0.25	1	4.514%	97	1	1	0.25	4.527%
8	0.75	0.5	0.5	4.491%	53	0	0.25	0.75	4.514%	98	0.5	0.25	0.5	4.529%
9	0.75	0.5	1	4.493%	54	0.25	0.25	1	4.514%	99	1	0.75	0.5	4.532%
10	1	1	1	4.493%	55	0.25	1	0.75	4.515%	100	0	1	0.75	4.532%
11	0.5	0.5	1	4.496%	56	0.25	0.75	1	4.515%	101	0.75	0.5	0.75	4.532%
12	0.5	0	0	4.497%	57	0	0.75	0	4.515%	102	1	0.25	0.5	4.532%
13	0.25	0.5	0.25	4.499%	58	0.5	0.5	0.25	4.515%	103	0.25	0.5	0.75	4.535%
14	0.5	0	1	4.499%	59	0.25	0.75	0.25	4.515%	104	0.25	1	0	4.536%
15	0	1	0	4.502%	60	0.5	0.75	0.5	4.516%	105	0.75	0.25	0.75	4.536%
16	0.5	0	0.75	4.503%	61	0.25	0.5	0	4.516%	106	1	0.5	0	4.537%
17	1	1	0	4.503%	62	0.5	0.75	0	4.516%	107	1	0.5	0.5	4.538%
18	0	0.25	0.25	4.504%	63	0.25	1	0.25	4.516%	108	1	0.25	0	4.538%
19	0.75	1	0.5	4.504%	64	0.5	0.25	0.25	4.516%	109	1	0.25	0.75	4.541%
20	0.75	0.75	0.75	4.505%	65	0.25	1	1	4.516%	110	1	0	0.75	4.548%
21	0.5	0.5	0.75	4.505%	66	0.75	0.5	0	4.516%	111	1	0.25	0.25	4.555%
22	0	0.5	0.5	4.505%	67	0.5	0.75	0.75	4.516%	112	1	0	1	4.557%
23	1	1	0.75	4.505%	68	0.5	0	0.5	4.516%	113	1	0	0.25	4.558%
24	0	0.5	0.25	4.506%	69	0.5	1	0.25	4.516%	114	1	0	0	4.560%
25	0	1	0.25	4.506%	70	0.75	0.25	0	4.517%	115	1	0	0.5	4.562%
26	0	1	0.5	4.506%	71	0.25	0.25	0	4.517%	116	0.25	0	1	4.581%
27	1	0.75	0.75	4.506%	72	0.75	0.25	0.25	4.517%	117	0.25	0	0	4.581%
28	0	0.5	1	4.507%	73	0.25	1	0.5	4.517%	118	0.25	0	0.75	4.586%
29	0	0.5	0	4.507%	74	0.25	0.5	0.5	4.517%	119	0.25	0	0.25	4.590%
30	0.75	0.75	1	4.508%	75	1	1	0.5	4.518%	120	0.25	0	0.5	4.595%
31	0.25	0.75	0	4.508%	76	0.5	0.25	0.75	4.519%	121	0	0	0	4.764%
32	0.25	0.5	1	4.509%	77	0	0.75	1	4.519%	122	0	0	0.5	4.769%
33	0.25	0.75	0.75	4.509%	78	0.5	0.75	1	4.519%	123	0	0	1	4.776%
34	0.75	1	0.75	4.509%	79	0	1	1	4.519%	124	0	0	0.25	4.776%
35	0.75	0.75	0.25	4.510%	80	0	0.25	0.5	4.520%	125	0	0	0.75	4.777%
36	0.25	0.25	0.75	4.510%	81	0.75	0.5	0.25	4.520%					
37	0.25	0.75	0.5	4.510%	82	0	0.25	0	4.522%					
38	0.5	0.75	0.25	4.511%	83	1	0.5	0.75	4.522%					
39	1	0.5	0.25	4.512%	84	1	0.25	1	4.522%					
40	1	0.75	0	4.512%	85	1	0.75	1	4.523%					
41	0.75	1	0	4.512%	86	0.25	0.25	0.5	4.523%					
42	0	0.5	0.75	4.512%	87	0	0.75	0.5	4.524%					
43	0.75	0.75	0	4.512%	88	0.5	1	0.75	4.524%					
44	0.5	0.5	0.5	4.512%	89	0.75	0.75	0.5	4.525%					
45	0	0.75	0.75	4.512%	90	0.5	1	0.5	4.525%					

Table C.6: Average monthly MAPE, model FreqLCA.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0.75	0	0	4.725%	46	0.75	0.25	0	5.012%	91	0.5	1	0.5	5.050%
2	0.75	0	0.5	4.734%	47	0.75	0.25	0.5	5.014%	92	1	1	0.75	5.050%
3	0.75	0	0.25	4.744%	48	0.25	0.75	1	5.018%	93	1	0	0.25	5.050%
4	0.75	0	0.75	4.747%	49	0.25	0.5	1	5.019%	94	1	0.5	0.5	5.050%
5	0.75	0	1	4.751%	50	0.5	0.5	0.5	5.022%	95	1	0	1	5.051%
6	0.5	0	0.25	4.751%	51	0.25	0.5	0.75	5.022%	96	0.75	1	1	5.052%
7	0.25	0	0.5	4.762%	52	0.5	0.5	0.75	5.023%	97	0.75	1	0	5.052%
8	0	0	0.75	4.764%	53	0.5	0.75	0	5.023%	98	1	1	0.5	5.053%
9	0.5	0	0	4.765%	54	0.5	0.5	0	5.024%	99	0	1	1	5.053%
10	0.25	0	0.75	4.768%	55	0	0.75	0.25	5.025%	100	1	0.75	0.5	5.053%
11	0.5	0	0.5	4.768%	56	0.75	0.25	0.25	5.026%	101	1	0.75	0.25	5.053%
12	0.5	0	0.75	4.774%	57	0.25	0.75	0.25	5.026%	102	0	1	0.5	5.054%
13	0.25	0	0	4.775%	58	0.75	0.5	0	5.027%	103	1	0.5	0.75	5.055%
14	0	0	0	4.781%	59	0.75	0.25	0.75	5.028%	104	0	1	0.75	5.057%
15	0.25	0	1	4.782%	60	0	0.5	0.75	5.028%	105	0.25	1	1	5.058%
16	0.5	0	1	4.783%	61	0.75	1	0.25	5.030%	106	0.5	1	0	5.058%
17	0	0	0.25	4.784%	62	0	0.75	1	5.030%	107	1	0.25	0.75	5.059%
18	0.25	0	0.25	4.787%	63	0.75	0.5	0.25	5.033%	108	0.5	1	1	5.060%
19	0	0	0.5	4.788%	64	1	0.25	0.25	5.033%	109	1	0.25	1	5.061%
20	0	0	1	4.792%	65	0.5	0.75	0.5	5.033%	110	0.75	0.75	0	5.061%
21	0	0.25	0	4.914%	66	1	0	0.5	5.035%	111	0.25	0.75	0.5	5.062%
22	0.25	0.25	0.75	4.937%	67	0.25	0.75	0.75	5.037%	112	1	0.75	0.75	5.062%
23	0	0.25	0.75	4.943%	68	0.75	0.75	0.75	5.037%	113	1	0.25	0.5	5.062%
24	0	0.25	0.5	4.944%	69	0	0.75	0.5	5.037%	114	0.75	0.75	1	5.062%
25	0.25	0.25	0.25	4.948%	70	0.75	0.5	1	5.037%	115	0.75	1	0.75	5.062%
26	0.25	0.25	0	4.949%	71	0.75	0.5	0.75	5.038%	116	0	1	0	5.063%
27	0	0.25	0.25	4.949%	72	0.5	1	0.25	5.038%	117	1	0.75	0	5.063%
28	0.25	0.25	1	4.953%	73	0.5	0.75	0.25	5.038%	118	1	0.25	0	5.064%
29	0	0.25	1	4.953%	74	1	0	0.75	5.038%	119	0.25	1	0.5	5.064%
30	0.25	0.25	0.5	4.955%	75	1	1	0	5.040%	120	1	1	1	5.066%
31	0.5	0.25	0	4.968%	76	0.75	0.75	0.25	5.040%	121	0.25	1	0	5.066%
32	0.5	0.25	1	4.973%	77	0	0.75	0.75	5.041%	122	0.5	1	0.75	5.066%
33	0.5	0.25	0.5	4.975%	78	1	0.5	1	5.041%	123	1	0.5	0	5.069%
34	0.5	0.25	0.25	4.984%	79	1	1	0.25	5.042%	124	1	0.75	1	5.074%
35	0.5	0.25	0.75	4.989%	80	0.5	0.75	1	5.042%	125	1	0	0	5.075%
36	0	0.5	0.25	4.997%	81	0.25	1	0.75	5.043%					
37	0.5	0.5	0.25	5.000%	82	0.75	1	0.5	5.043%					
38	0.25	0.5	0	5.001%	83	0	1	0.25	5.043%					
39	0	0.5	1	5.001%	84	0.75	0.5	0.5	5.043%					
40	0.25	0.5	0.25	5.005%	85	0.75	0.75	0.5	5.045%					
41	0	0.5	0	5.006%	86	0.5	0.75	0.75	5.045%					
42	0.25	0.5	0.5	5.006%	87	1	0.5	0.25	5.046%					
43	0.5	0.5	1	5.008%	88	0	0.75	0	5.046%					
44	0	0.5	0.5	5.010%	89	0.25	1	0.25	5.046%					
45	0.75	0.25	1	5.011%	90	0.25	0.75	0	5.047%					

Table C.7: Average monthly MAPE, model FreqMax.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0.75	0	1	4.392%	46	0.5	0.75	0.5	4.418%	91	1	0	1	4.431%
2	1	0.75	0.5	4.395%	47	0.25	0.25	1	4.418%	92	0.5	0.25	1	4.432%
3	0.75	0	0.75	4.395%	48	0.75	0.75	0.75	4.418%	93	1	0.25	0.25	4.432%
4	1	0.75	0.25	4.399%	49	0.5	0.75	0.25	4.418%	94	1	0.25	1	4.432%
5	0.5	0.75	1	4.399%	50	0.25	1	1	4.418%	95	0.5	0.25	0.75	4.432%
6	1	0.5	0.25	4.401%	51	0	0.5	0	4.418%	96	0.75	0.5	1	4.433%
7	0.75	0	0.5	4.403%	52	1	1	1	4.419%	97	0.25	0.5	0.25	4.433%
8	0.5	0.75	0	4.404%	53	0	0.25	1	4.419%	98	0.5	0.5	1	4.434%
9	0.75	0	0	4.404%	54	0.5	1	0.25	4.419%	99	1	0.75	0	4.434%
10	0	1	0.25	4.404%	55	0.75	0.75	0	4.419%	100	0.5	1	0.5	4.434%
11	1	0.25	0.5	4.405%	56	0.75	0.75	1	4.420%	101	1	0	0.25	4.434%
12	0.5	0.5	0.25	4.407%	57	0.25	0.5	0	4.421%	102	0.25	1	0.75	4.436%
13	0.5	0.5	0.5	4.409%	58	1	0	0.75	4.421%	103	0	0.25	0	4.436%
14	0.5	0.75	0.75	4.409%	59	0	0.75	0.25	4.421%	104	1	1	0	4.436%
15	0	1	0.5	4.409%	60	0	1	0.75	4.422%	105	0.5	0.25	0.5	4.436%
16	1	0.75	1	4.410%	61	1	0.5	0.75	4.422%	106	0	1	1	4.438%
17	1	0.5	0	4.410%	62	0.25	0.25	0	4.422%	107	0.5	0.25	0.25	4.439%
18	0	0.75	0.75	4.410%	63	0.5	1	0.75	4.422%	108	0.5	0	0	4.444%
19	0	0.5	0.25	4.410%	64	0.25	0.75	0	4.423%	109	1	1	0.75	4.445%
20	0	0.5	0.75	4.410%	65	0.75	0.5	0.5	4.423%	110	0.75	1	0.25	4.450%
21	0	0.75	0.5	4.410%	66	0.25	0.25	0.25	4.423%	111	0.25	0.75	1	4.450%
22	0	0.25	0.5	4.410%	67	0.25	1	0	4.424%	112	0.5	0	1	4.458%
23	0.75	1	0	4.411%	68	0.75	0.75	0.5	4.424%	113	0.5	0	0.5	4.467%
24	0.25	0.75	0.5	4.411%	69	0	0.5	0.5	4.425%	114	0.5	0	0.75	4.469%
25	1	0.5	0.5	4.411%	70	0	0.5	1	4.425%	115	0.5	0	0.25	4.472%
26	0.75	0.25	0.25	4.412%	71	0.75	0.5	0.25	4.425%	116	0.25	0	0.5	4.580%
27	0.25	0.25	0.75	4.413%	72	0.75	0.25	0.5	4.425%	117	0.25	0	0.75	4.581%
28	0.75	0.25	0.75	4.413%	73	0.25	0.75	0.25	4.425%	118	0.25	0	0	4.583%
29	0.25	0.25	0.5	4.413%	74	0.75	0	0.25	4.425%	119	0.25	0	1	4.584%
30	1	0.25	0.75	4.413%	75	1	0	0.5	4.425%	120	0.25	0	0.25	4.584%
31	1	0.5	1	4.414%	76	1	0.75	0.75	4.426%	121	0	0	0	4.767%
32	0.75	0.25	0	4.414%	77	0.25	0.5	0.75	4.426%	122	0	0	1	4.779%
33	0.75	1	0.75	4.415%	78	0	1	0	4.426%	123	0	0	0.75	4.785%
34	1	0	0	4.415%	79	0.75	0.5	0	4.426%	124	0	0	0.25	4.786%
35	1	1	0.25	4.415%	80	0.25	0.5	1	4.427%	125	0	0	0.5	4.788%
36	0.25	0.5	0.5	4.416%	81	0.75	0.75	0.25	4.427%					
37	0	0.75	0	4.416%	82	0.25	1	0.5	4.428%					
38	0	0.75	1	4.416%	83	0.25	0.75	0.75	4.428%					
39	0	0.25	0.75	4.416%	84	0.5	1	1	4.428%					
40	0.5	0.5	0.75	4.417%	85	1	0.25	0	4.428%					
41	0.25	1	0.25	4.417%	86	0.75	1	0.5	4.429%					
42	0.5	0.5	0	4.417%	87	1	1	0.5	4.429%					
43	0	0.25	0.25	4.417%	88	0.75	0.5	0.75	4.429%					
44	0.75	0.25	1	4.417%	89	0.5	1	0	4.430%					
45	0.5	0.25	0	4.417%	90	0.75	1	1	4.430%					

Table C.8: Average monthly MAPE, model FreqLogit.

Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE	Rank	c_π	c_μ	c_ν	MAPE
1	0.25	0.5	0.25	4.362%	46	0.75	0	0.25	4.396%	91	0.25	0.5	0.75	4.408%
2	0.75	0	0.75	4.365%	47	0	1	0.5	4.396%	92	0.5	0.75	0.75	4.408%
3	0.75	0	1	4.374%	48	0.25	0.5	0	4.398%	93	1	0.5	0.5	4.409%
4	0.75	0.75	0.5	4.375%	49	1	1	1	4.398%	94	0.5	1	0	4.409%
5	0.75	0	0	4.379%	50	0.25	1	0.5	4.398%	95	0.5	0.75	0.5	4.409%
6	0.25	1	0.25	4.381%	51	1	0	0	4.398%	96	0.75	0.25	1	4.410%
7	1	1	0.25	4.381%	52	0.5	0.25	0	4.399%	97	0	0.25	0	4.410%
8	0	1	0	4.382%	53	0.25	0.75	0	4.399%	98	0.75	0.75	1	4.410%
9	0.5	1	0.75	4.383%	54	0	1	0.75	4.400%	99	1	0.75	0.75	4.410%
10	0	0.75	0.5	4.383%	55	1	0.5	0.25	4.400%	100	0.5	0.75	1	4.413%
11	0.25	1	0	4.384%	56	0.75	0.5	0.5	4.400%	101	0.25	0.25	0	4.413%
12	0	0.5	0.75	4.386%	57	0.25	0.5	0.5	4.400%	102	0.25	1	1	4.413%
13	0	0.25	1	4.386%	58	0.75	0.5	1	4.401%	103	0.75	0.75	0.75	4.414%
14	0	0.25	0.5	4.387%	59	0.5	1	0.5	4.401%	104	0.5	0.5	0.25	4.414%
15	0.5	0.75	0.25	4.388%	60	0.5	0.25	1	4.401%	105	0.75	1	0	4.416%
16	0.5	1	0.25	4.388%	61	0.75	0.25	0	4.401%	106	0.25	0.75	0.5	4.419%
17	0.75	0.5	0.75	4.389%	62	0	1	1	4.401%	107	0.25	0.75	0.75	4.419%
18	0.75	0	0.5	4.389%	63	1	0.25	0.25	4.402%	108	0.25	0.25	0.75	4.420%
19	0.75	0.75	0	4.390%	64	0.25	0.5	1	4.402%	109	0.5	0.5	1	4.424%
20	0.5	0.5	0	4.390%	65	1	0.75	0.25	4.402%	110	0.5	0.25	0.75	4.424%
21	0	0.75	0.75	4.390%	66	0.5	0.75	0	4.402%	111	0.5	0	1	4.455%
22	0.75	0.75	0.25	4.390%	67	0.25	0.25	0.25	4.402%	112	0.5	0	0.75	4.456%
23	0	0.25	0.25	4.391%	68	0.5	0.5	0.5	4.403%	113	0.5	0	0.25	4.458%
24	1	0.75	1	4.391%	69	1	0.75	0	4.403%	114	0.5	0	0.5	4.478%
25	1	0	0.75	4.391%	70	0	0.5	0.25	4.404%	115	0.5	0	0	4.482%
26	0.25	0.25	0.5	4.391%	71	0.75	0.5	0	4.404%	116	0.25	0	0.75	4.572%
27	1	1	0.5	4.391%	72	1	0.75	0.5	4.404%	117	0.25	0	0.25	4.578%
28	0.5	0.25	0.25	4.391%	73	1	1	0.75	4.404%	118	0.25	0	1	4.580%
29	1	0	0.25	4.392%	74	0.5	0.25	0.5	4.405%	119	0.25	0	0.5	4.582%
30	1	0.5	1	4.392%	75	0	0.5	0	4.405%	120	0.25	0	0	4.582%
31	0.75	1	0.5	4.392%	76	1	0.25	0.75	4.405%	121	0	0	1	4.774%
32	0.75	0.25	0.5	4.392%	77	0.5	0.5	0.75	4.405%	122	0	0	0.5	4.777%
33	0.25	0.25	1	4.393%	78	0	0.5	0.5	4.405%	123	0	0	0.25	4.782%
34	1	1	0	4.393%	79	0.75	0.25	0.75	4.405%	124	0	0	0	4.786%
35	0.75	0.5	0.25	4.393%	80	0	0.25	0.75	4.405%	125	0	0	0.75	4.793%
36	1	0.25	1	4.394%	81	1	0.25	0	4.405%					
37	1	0.25	0.5	4.395%	82	1	0.5	0.75	4.405%					
38	1	0	0.5	4.395%	83	1	0.5	0	4.406%					
39	0.75	1	1	4.395%	84	0	0.5	1	4.406%					
40	1	0	1	4.395%	85	0.25	0.75	0.25	4.406%					
41	0.75	1	0.25	4.395%	86	0.25	1	0.75	4.407%					
42	0.75	1	0.75	4.396%	87	0	0.75	1	4.407%					
43	0.75	0.25	0.25	4.396%	88	0	0.75	0	4.407%					
44	0	0.75	0.25	4.396%	89	0.5	1	1	4.407%					
45	0.25	0.75	1	4.396%	90	0	1	0.25	4.408%					

Table C.9: Average monthly MAPE, model FreqRandom.

Appendix D

Graph generation results

Graph topology	All Previous			Creator			Last Post		
Time scale	Weekly	Monthly	Global	Weekly	Monthly	Global	Weekly	Monthly	Global
FreqLCA	p. 119	p. 123	p. 127	p. 131	p. 135	p. 139	p. 143	p. 147	p. 151
FreqMax	p. 120	p. 124	p. 128	p. 132	p. 136	p. 140	p. 144	p. 148	p. 152
FreqLogit	p. 121	p. 125	p. 129	p. 133	p. 137	p. 141	p. 145	p. 149	p. 153
FreqRandom	p. 122	p. 126	p. 130	p. 134	p. 138	p. 142	p. 146	p. 150	p. 154

Table D.1: F-measure results index.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.75	0.75	2.722%	46	1	0.5	1	2.693%	91	1	0.25	0.25	2.676%
2	0.75	0	0.5	2.722%	47	0	0	1	2.693%	92	0	0.75	0.5	2.676%
3	0.5	0.75	0.25	2.720%	48	1	0.75	0.75	2.692%	93	0	0	0	2.676%
4	0.75	0.25	0	2.718%	49	0.5	0.5	0.5	2.692%	94	1	0.25	0.75	2.675%
5	0	0.5	0	2.714%	50	0.75	1	0	2.692%	95	0.75	1	0.75	2.675%
6	0.5	1	0.5	2.714%	51	0.75	0.5	0.75	2.691%	96	1	0.75	0.5	2.675%
7	0	0.75	1	2.714%	52	0	0.25	1	2.691%	97	0.25	0.25	1	2.673%
8	0	1	1	2.713%	53	0.25	0.5	0.5	2.691%	98	0.25	0	0.5	2.672%
9	0.5	0.25	0.75	2.713%	54	1	0	0.5	2.690%	99	0.25	0	0.75	2.672%
10	1	0.25	1	2.713%	55	0.75	0	0	2.690%	100	0.25	0.25	0.5	2.671%
11	0.5	0	0.25	2.712%	56	0.5	0.5	0	2.689%	101	1	0	0.75	2.671%
12	0.75	0.25	0.75	2.711%	57	1	0	1	2.689%	102	0.75	0.5	1	2.671%
13	0	0.75	0.25	2.711%	58	0.75	0.25	0.5	2.689%	103	0.75	0.25	0.25	2.671%
14	0.5	0.75	0.5	2.711%	59	0	1	0.5	2.687%	104	0.5	1	1	2.670%
15	0.5	1	0.25	2.710%	60	0.5	0.75	0.75	2.687%	105	0	0.5	0.5	2.670%
16	0	0.25	0.5	2.710%	61	0.25	0.5	0	2.685%	106	0	1	0.25	2.669%
17	0.75	0.5	0.5	2.709%	62	0.25	1	1	2.685%	107	0.75	1	1	2.669%
18	0.75	0	0.75	2.708%	63	0.25	0.25	0.25	2.685%	108	0.25	0.5	0.25	2.669%
19	0.25	0.75	0.5	2.707%	64	1	0.5	0.25	2.684%	109	0.75	0.75	0.25	2.668%
20	0.75	0.75	0.5	2.707%	65	0.5	0.75	0	2.683%	110	0	1	0.75	2.665%
21	0.75	0.5	0.25	2.706%	66	1	0	0.25	2.683%	111	0.5	0.5	0.25	2.664%
22	0.5	0.25	0.25	2.706%	67	0.5	0.75	1	2.683%	112	0.5	0.25	0.5	2.664%
23	0.5	1	0	2.705%	68	0	1	0	2.683%	113	0	0	0.5	2.663%
24	0.75	0.75	1	2.705%	69	1	0.75	1	2.682%	114	0.25	0.75	1	2.663%
25	0.25	1	0	2.705%	70	0.25	0.5	0.75	2.682%	115	0.75	1	0.5	2.662%
26	0	0.5	0.25	2.704%	71	1	0.25	0	2.681%	116	0	0.25	0.75	2.661%
27	1	0.5	0	2.703%	72	1	1	0.75	2.681%	117	0.25	0.5	1	2.661%
28	0.25	0.25	0	2.703%	73	1	1	0.5	2.681%	118	0.25	1	0.25	2.661%
29	0.5	0	0.5	2.702%	74	0.25	0	0	2.681%	119	0.25	0.75	0.75	2.658%
30	1	0.5	0.5	2.702%	75	0.5	0.5	1	2.680%	120	0	0.75	0	2.658%
31	0.5	0	0	2.701%	76	1	0.25	0.5	2.680%	121	0	0.75	0.75	2.657%
32	1	0.75	0.25	2.701%	77	0.75	0	0.25	2.680%	122	0	0	0.75	2.657%
33	0.5	0.25	0	2.701%	78	0.75	0.5	0	2.680%	123	0.75	1	0.25	2.654%
34	1	1	0	2.701%	79	0.25	0.75	0	2.680%	124	1	0.5	0.75	2.652%
35	0	0.5	1	2.699%	80	0.5	0.25	1	2.679%	125	0.75	0	1	2.651%
36	0.5	0	0.75	2.699%	81	0.25	0.25	0.75	2.679%					
37	0.75	0.25	1	2.699%	82	0.25	0	1	2.679%					
38	0	0.5	0.75	2.699%	83	0.5	0.5	0.75	2.679%					
39	0	0.25	0	2.699%	84	0.25	1	0.5	2.679%					
40	1	1	0.25	2.698%	85	0.25	0	0.25	2.678%					
41	0.25	0.75	0.25	2.697%	86	0.5	0	1	2.678%					
42	0	0.25	0.25	2.697%	87	1	0	0	2.677%					
43	0	0	0.25	2.696%	88	0.75	0.75	0	2.677%					
44	0.25	1	0.75	2.695%	89	1	0.75	0	2.676%					
45	1	1	1	2.693%	90	0.5	1	0.75	2.676%					

Table D.2: F-measure of the FreqLCA model on a weekly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.25	0.25	5.577%	46	0.75	0	0.25	5.489%	91	0.25	1	0.75	5.441%
2	1	0.75	0.25	5.571%	47	0.25	1	0.5	5.489%	92	0	0.75	0.75	5.438%
3	0	0.75	0.5	5.550%	48	1	0.25	0.5	5.488%	93	1	1	0.75	5.436%
4	0.5	0	0	5.536%	49	0.75	1	0	5.486%	94	0.75	0.25	0.5	5.435%
5	1	0.25	0	5.535%	50	0.5	1	0.5	5.485%	95	0.25	0.75	0.75	5.434%
6	0	1	0.75	5.534%	51	1	0.25	0.25	5.485%	96	0.75	0.75	0.5	5.433%
7	0.25	1	0.25	5.532%	52	1	0	0.75	5.484%	97	0.75	0.25	0.75	5.431%
8	1	0	0.5	5.531%	53	0.5	0	0.5	5.484%	98	0.25	1	1	5.430%
9	0.5	0.5	0.25	5.531%	54	0.5	0.25	0	5.484%	99	0.5	1	1	5.430%
10	0.25	0	1	5.528%	55	0.5	0.75	0.25	5.483%	100	0	0.5	1	5.425%
11	0.25	0.25	0	5.527%	56	1	0.75	0.5	5.483%	101	0.75	1	0.75	5.425%
12	0	0	0.5	5.522%	57	1	1	0.5	5.481%	102	0.75	0.75	0.75	5.424%
13	1	0	0	5.521%	58	0.75	0.5	0	5.480%	103	0.25	0.5	0.75	5.424%
14	0.5	0	0.75	5.521%	59	0.25	0	0	5.478%	104	0.25	0.5	0.5	5.422%
15	0	0.75	0	5.518%	60	0.75	1	0.5	5.478%	105	0.25	0.75	0.5	5.419%
16	0	0	0.25	5.518%	61	0.5	0.25	0.75	5.478%	106	0	0.25	0.75	5.412%
17	0	0.5	0	5.517%	62	0	0.25	0	5.477%	107	0.5	1	0.75	5.412%
18	0.25	1	0	5.517%	63	0.25	0.75	0	5.476%	108	0	0.5	0.75	5.411%
19	0.75	1	0.25	5.516%	64	0.75	0	0.75	5.476%	109	0	0.25	1	5.408%
20	0.5	0.75	0.75	5.514%	65	0.75	0.5	0.5	5.474%	110	0.75	0.5	1	5.406%
21	0	0.5	0.5	5.514%	66	0.75	0	0	5.473%	111	0.75	0.75	1	5.404%
22	0.75	0.5	0.25	5.513%	67	1	0	0.25	5.473%	112	0.5	0.5	1	5.397%
23	1	1	0.25	5.513%	68	0.5	0.25	0.25	5.472%	113	1	0.75	1	5.396%
24	0	1	0.25	5.512%	69	0.25	0.25	0.5	5.472%	114	0.25	0.75	1	5.384%
25	0.5	0.75	0	5.510%	70	1	0.5	0	5.471%	115	0.5	0.5	0.75	5.378%
26	0.25	0.75	0.25	5.510%	71	0.25	0.5	0	5.469%	116	0.75	0.25	1	5.377%
27	0	0.5	0.25	5.507%	72	1	0.5	0.75	5.469%	117	0.25	0.25	1	5.375%
28	1	0	1	5.505%	73	0.5	0.25	0.5	5.468%	118	0.75	1	1	5.371%
29	0.5	0.5	0.5	5.505%	74	0.75	0	1	5.466%	119	0	1	1	5.371%
30	1	0.5	0.25	5.504%	75	1	0.75	0.75	5.462%	120	1	1	1	5.371%
31	0	0.25	0.25	5.503%	76	0.5	0.5	0	5.458%	121	0.5	0.25	1	5.369%
32	0.25	0	0.5	5.502%	77	1	0.75	0	5.458%	122	1	0.25	1	5.357%
33	0.5	0.75	0.5	5.502%	78	0.25	0	0.75	5.458%	123	0	0.75	1	5.352%
34	0.25	0.25	0.25	5.499%	79	0.25	0.25	0.75	5.457%	124	1	0.5	1	5.331%
35	0.5	1	0	5.498%	80	0	1	0.5	5.455%	125	0.25	0.5	1	5.310%
36	0.5	0	1	5.496%	81	1	0.25	0.75	5.455%					
37	0	0	0.75	5.496%	82	0	1	0	5.454%					
38	0.75	0.25	0	5.495%	83	0.75	0.75	0.25	5.454%					
39	0.5	1	0.25	5.493%	84	0.75	0.75	0	5.452%					
40	1	1	0	5.493%	85	0	0	0	5.451%					
41	0.5	0	0.25	5.492%	86	0	0.25	0.5	5.451%					
42	0	0.75	0.25	5.490%	87	1	0.5	0.5	5.450%					
43	0.25	0	0.25	5.489%	88	0.75	0.5	0.75	5.444%					
44	0.25	0.5	0.25	5.489%	89	0.5	0.75	1	5.442%					
45	0	0	1	5.489%	90	0.75	0	0.5	5.441%					

Table D.3: F-measure of the FreqMax model on a weekly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	1	1	2.764%	46	1	0.25	0.75	2.709%	91	0.5	0.5	0	2.694%
2	0.75	0	0.75	2.746%	47	0.75	0.75	1	2.709%	92	0.75	0.75	0.75	2.694%
3	0.25	1	1	2.742%	48	0.5	0.5	0.75	2.708%	93	0.5	0	1	2.692%
4	0.5	0	0	2.737%	49	0.5	0.75	0.5	2.707%	94	1	0.5	0.25	2.692%
5	0.25	1	0.25	2.736%	50	0	1	0.25	2.707%	95	0	0.25	0.5	2.691%
6	0.75	0.5	0.75	2.736%	51	1	0.5	0.5	2.707%	96	0.5	1	0	2.691%
7	0	1	0.75	2.736%	52	0.75	0.25	1	2.706%	97	0.25	0.5	0	2.691%
8	0	0.5	0.25	2.736%	53	0.75	0.75	0.25	2.705%	98	1	0.5	1	2.690%
9	0.25	0	0.5	2.730%	54	0.5	0.5	0.25	2.705%	99	0.5	0	0.75	2.689%
10	0.5	0.25	1	2.729%	55	0.25	0.5	1	2.705%	100	0	0.75	1	2.689%
11	0.25	0.25	0.5	2.729%	56	0	0	0.75	2.704%	101	0.75	1	0.75	2.689%
12	0	1	1	2.728%	57	0	0.75	0.25	2.704%	102	0.5	1	0.5	2.688%
13	0	0.5	0.5	2.726%	58	0.75	0	0.25	2.703%	103	0.75	0	1	2.687%
14	1	0.25	0	2.726%	59	0.75	1	0.5	2.703%	104	0	1	0	2.687%
15	0	0	0.25	2.726%	60	0.5	0.25	0	2.703%	105	0.5	0.25	0.5	2.687%
16	0	0	0.5	2.724%	61	0	0.75	0.75	2.702%	106	1	0.75	0.5	2.687%
17	0.25	0	0.25	2.724%	62	1	1	0.75	2.702%	107	0.75	0.75	0.5	2.686%
18	0.25	0	0.75	2.723%	63	0.25	1	0.5	2.702%	108	0.5	1	1	2.684%
19	0.25	0	0	2.723%	64	0	0	0	2.702%	109	1	0	0.75	2.684%
20	1	0.25	0.5	2.722%	65	0.25	0.25	1	2.702%	110	1	0	0.5	2.684%
21	1	0.25	1	2.722%	66	0	0.25	0	2.701%	111	0.75	0.25	0.5	2.683%
22	0.25	0.5	0.75	2.722%	67	0.5	0.75	0	2.701%	112	1	0.5	0.75	2.681%
23	0	1	0.5	2.721%	68	0.25	0.75	0	2.700%	113	0.5	1	0.25	2.680%
24	0.75	0.5	1	2.721%	69	0.25	0.75	0.75	2.700%	114	0.25	0	1	2.680%
25	0.25	0.25	0.75	2.721%	70	1	0	1	2.700%	115	1	1	0.25	2.679%
26	0	0.5	1	2.720%	71	1	0.75	0	2.699%	116	0.5	0	0.5	2.679%
27	0.75	1	0	2.719%	72	0	0.5	0.75	2.699%	117	0.5	0.75	0.75	2.679%
28	0.25	0.25	0.25	2.717%	73	0	0.75	0.5	2.699%	118	1	0	0.25	2.677%
29	1	0	0	2.717%	74	0.75	0	0	2.697%	119	0.5	0	0.25	2.676%
30	1	0.75	0.25	2.716%	75	0.5	0.75	1	2.697%	120	0.5	0.25	0.75	2.674%
31	0.75	0.25	0.75	2.716%	76	1	1	0	2.697%	121	0.75	0.5	0.5	2.672%
32	0.25	0.75	0.5	2.715%	77	1	0.25	0.25	2.697%	122	0.75	0.75	0	2.669%
33	1	0.75	1	2.714%	78	0.5	0.25	0.25	2.697%	123	0.25	1	0	2.665%
34	0.75	0.5	0	2.714%	79	0	0.5	0	2.696%	124	1	0.5	0	2.665%
35	0.25	0.75	0.25	2.714%	80	0.5	0.75	0.25	2.696%	125	0.5	0.5	1	2.664%
36	1	1	0.5	2.714%	81	1	0.75	0.75	2.696%					
37	0.25	0.5	0.25	2.713%	82	0	0.25	0.75	2.696%					
38	0.75	0.25	0	2.713%	83	0.75	1	1	2.695%					
39	0.25	1	0.75	2.713%	84	0.5	1	0.75	2.695%					
40	0.75	1	0.25	2.710%	85	0	0.25	1	2.695%					
41	0.25	0.5	0.5	2.710%	86	0.75	0.5	0.25	2.695%					
42	0.75	0	0.5	2.710%	87	0.5	0.5	0.5	2.694%					
43	0	0.25	0.25	2.710%	88	0.25	0.25	0	2.694%					
44	0	0	1	2.709%	89	0.75	0.25	0.25	2.694%					
45	0.25	0.75	1	2.709%	90	0	0.75	0	2.694%					

Table D.4: F-measure of the FreqLogit model on a weekly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.75	0.5	2.753%	46	1	0.5	0.5	2.709%	91	0.5	0.25	0.25	2.692%
2	1	0	0.5	2.741%	47	0	0.25	0.25	2.708%	92	0.5	0	0.75	2.692%
3	0.75	0	0	2.739%	48	0	0	0.25	2.708%	93	0.25	0.75	0.25	2.690%
4	0.5	1	0.75	2.736%	49	0.75	0.25	0	2.708%	94	1	0.75	0.5	2.690%
5	0.75	1	0.75	2.736%	50	0.75	0.75	0	2.707%	95	0	1	0.25	2.690%
6	0.5	0.75	0.5	2.736%	51	0.25	0.75	0.75	2.707%	96	0.25	0.5	0	2.690%
7	0.5	0.25	0.5	2.735%	52	0.25	0.5	0.5	2.707%	97	1	0.75	0	2.689%
8	1	0.75	0.75	2.734%	53	1	0.75	1	2.707%	98	1	0	0.25	2.689%
9	1	0.25	1	2.734%	54	0.5	0.5	0.5	2.707%	99	0.25	1	1	2.689%
10	0	0.75	1	2.729%	55	0.5	0.5	0	2.706%	100	0.5	0.75	0.25	2.688%
11	0.25	1	0.5	2.729%	56	1	0	0	2.706%	101	0.25	0.25	0	2.688%
12	1	1	0	2.728%	57	1	1	0.25	2.705%	102	0.5	0.75	1	2.687%
13	1	0.25	0	2.726%	58	0	0.25	1	2.705%	103	0.5	0.5	0.75	2.687%
14	1	0.25	0.25	2.725%	59	0	0.5	1	2.704%	104	0.25	1	0	2.687%
15	0	1	0.75	2.725%	60	0	0.5	0.5	2.704%	105	0	0.75	0.25	2.687%
16	0.25	0	1	2.724%	61	0	0.25	0.75	2.704%	106	0.25	0.5	0.75	2.686%
17	1	0.5	0.25	2.724%	62	0.25	0	0.75	2.703%	107	0.25	1	0.75	2.686%
18	1	0.5	0	2.723%	63	0.75	0.75	1	2.703%	108	0.75	0	1	2.685%
19	0.75	0.5	0.5	2.722%	64	0	1	0	2.703%	109	0.5	0	0.5	2.685%
20	0.75	0	0.5	2.722%	65	0.5	1	0.25	2.703%	110	0.25	0.5	1	2.684%
21	1	0.5	1	2.722%	66	1	1	0.75	2.702%	111	1	1	1	2.684%
22	0.25	0.75	1	2.721%	67	0.75	1	0.5	2.702%	112	0	0	0.5	2.683%
23	0	0.75	0.75	2.721%	68	0.75	0	0.25	2.702%	113	0	0.75	0	2.682%
24	0.75	1	0.25	2.720%	69	0.5	1	1	2.701%	114	0.25	0.25	0.5	2.681%
25	0.75	0.75	0.75	2.719%	70	0.25	0.25	1	2.701%	115	0	0.5	0.75	2.679%
26	0	0	0.75	2.719%	71	0.5	1	0.5	2.700%	116	0.5	0.25	0.75	2.678%
27	0.75	0.25	0.5	2.719%	72	0.5	0.75	0	2.700%	117	0	0	0	2.678%
28	0.75	0.5	0.75	2.719%	73	0	0.25	0	2.699%	118	0.75	0.25	0.25	2.677%
29	0.75	0.25	0.75	2.718%	74	0.25	0	0.25	2.699%	119	0.5	0.75	0.75	2.677%
30	0.25	0.5	0.25	2.718%	75	0.75	0.5	1	2.698%	120	0.5	1	0	2.675%
31	0.5	0.5	0.25	2.717%	76	0.25	0.25	0.25	2.698%	121	0	0.25	0.5	2.672%
32	0.25	0	0.5	2.717%	77	1	0	1	2.698%	122	0.25	0	0	2.672%
33	0	0.5	0	2.716%	78	0.5	0.25	0	2.697%	123	1	0.5	0.75	2.669%
34	0.75	0.75	0.25	2.716%	79	0	0.5	0.25	2.697%	124	0.5	0.5	1	2.665%
35	0.5	0.25	1	2.715%	80	0.75	0	0.75	2.697%	125	1	1	0.5	2.658%
36	0.75	0.25	1	2.714%	81	1	0.25	0.75	2.696%					
37	1	0	0.75	2.714%	82	0.25	0.25	0.75	2.696%					
38	0	0.75	0.5	2.713%	83	0.75	1	0	2.696%					
39	0.5	0	0	2.713%	84	1	0.25	0.5	2.695%					
40	0.25	1	0.25	2.712%	85	0	1	0.5	2.695%					
41	0.75	0.5	0.25	2.712%	86	0.25	0.75	0	2.694%					
42	0	0	1	2.711%	87	0	1	1	2.694%					
43	1	0.75	0.25	2.711%	88	0.5	0	1	2.694%					
44	0.25	0.75	0.5	2.711%	89	0.5	0	0.25	2.694%					
45	0.75	1	1	2.710%	90	0.75	0.5	0	2.693%					

Table D.5: F-measure of the FreqRandom model on a weekly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.75	0.75	7.413%	46	0	0.25	1	7.343%	91	0.75	0.25	0	7.314%
2	0.5	0	0.25	7.403%	47	1	0.75	0.5	7.343%	92	0.5	0.75	1	7.313%
3	0	0.5	0	7.399%	48	1	0.75	1	7.342%	93	0.25	0.25	0.25	7.313%
4	0.5	0	0	7.399%	49	0	0.75	0.25	7.342%	94	0.75	0.5	1	7.312%
5	1	0.5	0.5	7.395%	50	1	1	0.25	7.342%	95	1	0.75	0	7.312%
6	1	1	0	7.391%	51	0.75	1	1	7.339%	96	1	0.75	0.75	7.311%
7	0.75	0.5	0.25	7.390%	52	0.75	0.25	0.75	7.339%	97	0.25	0.25	0.75	7.310%
8	0	0.5	0.75	7.388%	53	0.75	1	0	7.337%	98	1	0.25	0.5	7.310%
9	0	0.5	1	7.385%	54	1	0.25	0.25	7.337%	99	0.75	0	0.25	7.309%
10	1	0	0.5	7.384%	55	1	0.75	0.25	7.337%	100	0.25	0.75	0.75	7.309%
11	0.5	1	0.25	7.383%	56	0.75	0.25	0.25	7.337%	101	0.25	0.75	1	7.307%
12	0.25	0.75	0.5	7.381%	57	0.5	0.75	0	7.337%	102	0.75	1	0.5	7.306%
13	0.75	0	0.5	7.381%	58	0.75	0	0	7.336%	103	0.25	0	0.5	7.305%
14	0	0.5	0.25	7.379%	59	0.25	0.75	0	7.336%	104	0.5	1	1	7.302%
15	0.5	0.75	0.5	7.375%	60	0.75	0.5	0.75	7.335%	105	0.25	1	0.25	7.300%
16	0	1	0	7.373%	61	0.75	0.5	0	7.334%	106	0.5	1	0.75	7.297%
17	0.5	0.5	0.5	7.370%	62	1	0	1	7.334%	107	0.75	0.25	1	7.297%
18	0.25	0	1	7.368%	63	0.5	0.5	0	7.334%	108	0	0.75	0.75	7.295%
19	0.25	1	0	7.368%	64	1	1	0.5	7.332%	109	0.25	0	0.75	7.293%
20	0.75	0.5	0.5	7.365%	65	0	0	0.25	7.332%	110	1	0.25	0	7.291%
21	0	1	1	7.365%	66	0.25	0.5	0.5	7.332%	111	0.5	0.25	1	7.290%
22	0.75	0.75	1	7.364%	67	0.25	0.5	0	7.330%	112	0	0.5	0.5	7.289%
23	0.75	0.75	0.5	7.363%	68	0.5	0.25	0.75	7.330%	113	1	0.5	1	7.289%
24	0.5	0.75	0.75	7.363%	69	0.25	0	0	7.328%	114	1	0	0.75	7.288%
25	0	0.25	0	7.362%	70	0.5	0.5	1	7.327%	115	0.5	0.25	0.5	7.284%
26	0	0.25	0.25	7.362%	71	0.25	0	0.25	7.327%	116	0.5	0.5	0.25	7.282%
27	1	1	1	7.362%	72	0.5	1	0	7.327%	117	0	0	0.5	7.279%
28	0.75	0.25	0.5	7.361%	73	0.75	0.75	0	7.325%	118	1	1	0.75	7.272%
29	0.25	0.25	1	7.360%	74	0.5	0.25	0.25	7.324%	119	0.25	0.5	1	7.271%
30	0	1	0.25	7.360%	75	0.5	0.75	0.25	7.324%	120	0.75	0	1	7.266%
31	1	0.5	0.25	7.359%	76	0.25	0.5	0.75	7.323%	121	0.75	1	0.75	7.264%
32	0.5	1	0.5	7.357%	77	0.5	0	0.5	7.322%	122	0	0	0.75	7.260%
33	0	0	1	7.353%	78	0.5	0.5	0.75	7.322%	123	1	0.5	0.75	7.255%
34	0	0.25	0.5	7.352%	79	1	0.5	0	7.322%	124	0	0.75	0	7.249%
35	0.25	0.75	0.25	7.351%	80	0	1	0.5	7.322%	125	0.75	1	0.25	7.246%
36	0	0	0	7.350%	81	0.5	0.25	0	7.322%					
37	1	0.25	1	7.349%	82	0.5	0	1	7.321%					
38	0.5	0	0.75	7.349%	83	0	1	0.75	7.321%					
39	0	0.75	1	7.348%	84	0	0.25	0.75	7.320%					
40	1	0	0	7.347%	85	0	0.75	0.5	7.320%					
41	0.75	0.75	0.25	7.346%	86	0.25	1	0.75	7.319%					
42	0.25	0.5	0.25	7.344%	87	0.25	0.25	0.5	7.319%					
43	0.25	0.25	0	7.343%	88	0.25	1	0.5	7.318%					
44	0.25	1	1	7.343%	89	0.75	0	0.75	7.316%					
45	1	0	0.25	7.343%	90	1	0.25	0.75	7.315%					

Table D.6: F-measure of the FreqLCA model on a monthly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.25	0.5	0.25	12.434%	46	0	0.5	0	12.368%	91	0.75	0.75	0.5	12.278%
2	0.5	0.5	0.25	12.433%	47	0.75	0.25	0	12.367%	92	0.25	0.75	0.75	12.275%
3	0.25	0	1	12.426%	48	0.75	0	0.5	12.364%	93	0	0.25	0.5	12.273%
4	0.25	1	0.25	12.424%	49	0.5	1	0	12.363%	94	0	0.25	0.75	12.273%
5	0.75	1	0.25	12.424%	50	0.5	0	0.25	12.363%	95	0.75	1	0.5	12.271%
6	0.75	0.5	0.25	12.419%	51	0	1	0.75	12.361%	96	0	0.5	0.75	12.269%
7	1	0	0.75	12.414%	52	0	0.25	0.25	12.361%	97	0.75	0.5	0.75	12.268%
8	1	0	0	12.413%	53	0.5	0.75	0	12.359%	98	0.5	0.5	1	12.266%
9	0.25	0	0.5	12.409%	54	0.75	0.75	0	12.359%	99	0.75	1	0.75	12.265%
10	0.25	0.5	0	12.409%	55	0.25	0	0.25	12.357%	100	0.25	1	1	12.257%
11	0.75	0.25	0.25	12.405%	56	1	1	0.25	12.357%	101	1	0.5	0.75	12.254%
12	0	0.75	0.5	12.402%	57	0	1	0	12.355%	102	0.25	0.5	0.75	12.254%
13	0	0.75	0.25	12.401%	58	0	0	0	12.354%	103	0	1	1	12.242%
14	1	0	0.5	12.401%	59	0.25	0	0.75	12.354%	104	0.75	0.75	0.75	12.235%
15	1	0	1	12.400%	60	0	1	0.5	12.352%	105	0	0.5	1	12.234%
16	1	0.75	0.25	12.399%	61	0.5	0.75	0.75	12.351%	106	1	1	0.75	12.234%
17	0	0.25	0	12.396%	62	0.5	0.5	0.5	12.350%	107	0.5	0.5	0.75	12.232%
18	0	0	0.75	12.394%	63	0.75	0.5	0	12.349%	108	1	0.25	0.75	12.229%
19	1	0.5	0.25	12.393%	64	0.5	0.75	0.5	12.348%	109	0.5	0.75	1	12.226%
20	1	0	0.25	12.391%	65	0.25	1	0	12.347%	110	0	0.25	1	12.197%
21	0.5	0.25	0.25	12.390%	66	1	0.25	0.25	12.347%	111	0.5	1	0.75	12.196%
22	0	0	1	12.390%	67	0	0.5	0.25	12.346%	112	0.75	0.5	1	12.185%
23	0	0	0.25	12.389%	68	0.5	0.25	0	12.343%	113	0.25	0.75	1	12.176%
24	0.5	0	0	12.388%	69	0.25	0.25	0	12.341%	114	1	0.75	1	12.176%
25	0	0.75	0	12.388%	70	1	0.75	0.75	12.338%	115	0.75	0.75	1	12.173%
26	0.25	0.75	0.25	12.387%	71	0.5	0	1	12.337%	116	0.25	0.25	1	12.168%
27	0.75	0	0	12.387%	72	0.5	0.25	0.5	12.334%	117	0.5	1	1	12.165%
28	0.75	0	0.25	12.386%	73	1	0.75	0	12.331%	118	1	0.25	1	12.148%
29	0.25	0.75	0	12.386%	74	0.25	1	0.75	12.327%	119	1	1	1	12.143%
30	0.75	0	0.75	12.384%	75	0.5	0.5	0	12.320%	120	0.25	0.5	1	12.142%
31	0.5	0	0.75	12.382%	76	0.25	1	0.5	12.314%	121	0	0.75	1	12.134%
32	1	1	0.5	12.382%	77	0.5	0	0.5	12.309%	122	0.75	0.25	1	12.128%
33	0.5	1	0.5	12.379%	78	0.25	0.5	0.5	12.308%	123	0.5	0.25	1	12.111%
34	1	0.5	0	12.379%	79	0.75	0.5	0.5	12.307%	124	0.75	1	1	12.098%
35	0	0	0.5	12.379%	80	0.75	0.75	0.25	12.305%	125	1	0.5	1	12.044%
36	1	0.75	0.5	12.378%	81	0.25	0.25	0.25	12.303%					
37	0.5	1	0.25	12.378%	82	1	0.25	0.5	12.302%					
38	0	1	0.25	12.377%	83	0	0.75	0.75	12.301%					
39	1	1	0	12.376%	84	0.5	0.25	0.75	12.292%					
40	0.75	1	0	12.376%	85	0.75	0.25	0.5	12.289%					
41	0.25	0	0	12.374%	86	0.75	0.25	0.75	12.289%					
42	0.5	0.75	0.25	12.372%	87	0.25	0.25	0.75	12.287%					
43	0.25	0.25	0.5	12.371%	88	1	0.5	0.5	12.286%					
44	1	0.25	0	12.369%	89	0.75	0	1	12.285%					
45	0	0.5	0.5	12.369%	90	0.25	0.75	0.5	12.283%					

Table D.7: F-measure of the FreqMax model on a monthly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	1	1	7.485%	46	0.75	0.25	1	7.403%	91	0.5	0.5	0.25	7.374%
2	0.25	1	0.25	7.463%	47	0.75	0.75	0.75	7.403%	92	0	0.5	0	7.373%
3	0.75	0.25	0.75	7.458%	48	0.25	0.25	1	7.402%	93	0	0.75	0	7.373%
4	0.75	0.75	1	7.457%	49	0.25	0.5	0.25	7.402%	94	1	0.5	1	7.373%
5	0	0.25	0.25	7.446%	50	0	0.75	0.25	7.401%	95	0.5	0	1	7.372%
6	0	0.5	0.25	7.445%	51	1	0.75	0	7.401%	96	0.25	0.75	0	7.371%
7	0.5	0.75	0	7.444%	52	0.75	0	1	7.400%	97	0	0.75	1	7.370%
8	0	0.5	1	7.444%	53	1	0.75	0.5	7.399%	98	1	0.25	0.25	7.370%
9	1	0.25	1	7.444%	54	0.25	1	1	7.399%	99	1	0	1	7.366%
10	0	0	0.25	7.441%	55	0.25	0.25	0.75	7.398%	100	0.5	1	0.25	7.366%
11	0.75	0	0.75	7.441%	56	0	0.25	1	7.396%	101	0.25	0.75	0.25	7.365%
12	0.75	1	0.5	7.438%	57	0.5	0.5	0	7.395%	102	0.25	0.5	1	7.365%
13	0.75	0	0.5	7.436%	58	0	1	0	7.394%	103	0.5	0	0.25	7.364%
14	0.5	0.75	0.5	7.436%	59	1	0	0	7.394%	104	0	0	0.5	7.364%
15	0.5	0	0	7.434%	60	0.75	0.75	0.25	7.393%	105	0.75	0.75	0	7.364%
16	0.75	0.5	1	7.433%	61	0.75	0.5	0.25	7.393%	106	0.5	0.75	0.75	7.363%
17	0.5	0.25	1	7.433%	62	0.5	0.75	1	7.393%	107	0.75	0.75	0.5	7.362%
18	0	0.25	0.75	7.433%	63	0.25	1	0.75	7.393%	108	0.5	1	0.75	7.361%
19	0.25	0	0.25	7.432%	64	0.75	0.25	0	7.393%	109	1	0.75	0.75	7.361%
20	0.25	0.5	0.75	7.432%	65	1	0.25	0.5	7.392%	110	0.25	0.5	0	7.361%
21	0.25	0.75	1	7.432%	66	0	1	0.5	7.390%	111	1	0.5	0.75	7.357%
22	0.75	0.25	0.25	7.431%	67	1	0	0.5	7.388%	112	1	0.25	0	7.356%
23	1	0.75	0.25	7.429%	68	0.5	1	1	7.386%	113	1	0	0.25	7.355%
24	1	0.75	1	7.425%	69	0	1	0.25	7.386%	114	1	1	0.5	7.354%
25	0.5	1	0	7.425%	70	0	0.75	0.5	7.385%	115	0.25	0.75	0.75	7.351%
26	0.5	0.5	0.75	7.424%	71	0.5	0.5	0.5	7.384%	116	0	0.75	0.75	7.351%
27	0.25	1	0	7.422%	72	0.5	0.25	0.75	7.383%	117	0.5	0.5	1	7.344%
28	0	1	1	7.421%	73	1	0.5	0.25	7.383%	118	0.75	0.5	0.5	7.344%
29	0.25	0.5	0.5	7.419%	74	0.25	0	1	7.381%	119	0.25	0.25	0.5	7.343%
30	0.75	1	0.25	7.419%	75	0.5	1	0.5	7.381%	120	1	0.25	0.75	7.339%
31	0.25	0.25	0.25	7.418%	76	0.75	0.25	0.5	7.380%	121	0	0	0	7.334%
32	0	0	0.75	7.416%	77	1	1	0	7.380%	122	0.5	0.25	0.5	7.332%
33	0	0.5	0.75	7.416%	78	0.75	0.5	0.75	7.380%	123	1	0	0.75	7.331%
34	0.25	0	0.75	7.416%	79	0	1	0.75	7.380%	124	0.75	1	0.75	7.331%
35	0.25	0	0	7.415%	80	0	0.25	0.5	7.379%	125	1	0.5	0	7.319%
36	0.25	1	0.5	7.415%	81	0.5	0.25	0.25	7.379%					
37	1	1	0.25	7.414%	82	0.75	1	0	7.379%					
38	0.25	0.75	0.5	7.413%	83	0.75	0	0.25	7.378%					
39	1	0.5	0.5	7.412%	84	0.5	0.75	0.25	7.378%					
40	0.25	0	0.5	7.411%	85	0	0.5	0.5	7.378%					
41	0	0	1	7.407%	86	0.5	0	0.5	7.377%					
42	1	1	0.75	7.407%	87	0.5	0.25	0	7.376%					
43	0.25	0.25	0	7.407%	88	0.75	1	1	7.376%					
44	0.75	0	0	7.406%	89	0.75	0.5	0	7.376%					
45	0	0.25	0	7.403%	90	0.5	0	0.75	7.374%					

Table D.8: F-measure of the FreqLogit model on a monthly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	1	0.75	7.473%	46	1	0.25	1	7.408%	91	0.5	0.5	0.5	7.382%
2	0.75	0.5	0.75	7.468%	47	0	0.5	0.5	7.408%	92	0	0.5	0.25	7.381%
3	1	0	0.5	7.467%	48	0.75	0.25	0	7.407%	93	0.75	1	1	7.381%
4	0.5	0	0	7.456%	49	0.75	0.5	0	7.407%	94	0.25	0.5	0.75	7.378%
5	0.25	0.75	0.5	7.454%	50	1	0	1	7.407%	95	1	0.75	0.5	7.378%
6	0.75	0.75	0.5	7.450%	51	1	0.5	0.75	7.405%	96	0.5	0.25	1	7.378%
7	0.25	0	1	7.449%	52	0.75	1	0.25	7.405%	97	0.5	0.75	1	7.378%
8	0.75	0.75	0.25	7.441%	53	0.25	0.5	0.25	7.405%	98	0.5	0.5	0.75	7.376%
9	0.5	0.75	0.5	7.440%	54	0.25	0	0.25	7.404%	99	0.5	1	0.75	7.376%
10	1	0.75	0.25	7.440%	55	0.75	1	0.5	7.404%	100	0.25	0	0	7.375%
11	0.5	0.25	0	7.439%	56	1	0	0	7.403%	101	0.75	0.25	1	7.375%
12	0.25	0	0.75	7.438%	57	0.75	0	0.25	7.403%	102	0.25	0.25	0.75	7.375%
13	0.75	0	0.5	7.437%	58	1	0.25	0	7.403%	103	1	0	0.25	7.374%
14	0.25	0.5	0.5	7.437%	59	0.25	1	1	7.403%	104	0	1	0	7.373%
15	0.25	1	0.75	7.437%	60	0	0	0	7.402%	105	0.5	1	0.25	7.370%
16	1	0.75	0.75	7.436%	61	1	0.75	0	7.401%	106	0.25	0.5	0	7.368%
17	1	1	0	7.433%	62	0.25	1	0.5	7.401%	107	0.5	0.25	0.75	7.367%
18	0.5	0.25	0.25	7.430%	63	0.75	0.5	0.5	7.400%	108	0.75	0	0.75	7.364%
19	1	0.5	1	7.430%	64	0.75	1	0	7.400%	109	0.5	0.5	1	7.363%
20	0	0.5	0	7.429%	65	0.5	0.75	0.75	7.400%	110	0.5	0	0.25	7.358%
21	0	0.5	1	7.427%	66	0.25	0.25	1	7.400%	111	0.5	0	1	7.358%
22	0	0.75	0.75	7.427%	67	0	0	1	7.399%	112	0.25	0.5	1	7.357%
23	0.75	0.75	0.75	7.426%	68	0	0	0.5	7.397%	113	0	0.75	0	7.351%
24	0	0.25	0	7.425%	69	0.75	0.25	0.5	7.397%	114	0.25	0.25	0	7.350%
25	0.75	0.5	0.25	7.424%	70	0.75	0.25	0.75	7.397%	115	1	0.75	1	7.350%
26	0.75	0.75	0	7.423%	71	0.5	1	0.5	7.394%	116	1	0.25	0.75	7.350%
27	0.5	0.25	0.5	7.421%	72	0.5	1	1	7.394%	117	0	1	0.25	7.348%
28	1	0.5	0.5	7.420%	73	0.25	1	0.25	7.394%	118	0.25	0.25	0.5	7.345%
29	0	0.75	0.5	7.420%	74	0	0.25	1	7.394%	119	0	0.75	0.25	7.339%
30	0.5	0.75	0	7.420%	75	0.5	0	0.5	7.390%	120	0	0.5	0.75	7.334%
31	1	0.5	0.25	7.419%	76	0.25	1	0	7.390%	121	0	1	1	7.333%
32	0	1	0.75	7.419%	77	1	1	0.75	7.389%	122	0	0.25	0.5	7.331%
33	0.75	0	0	7.417%	78	0.25	0.75	1	7.389%	123	0.75	0	1	7.330%
34	0.25	0	0.5	7.415%	79	0.25	0.75	0.25	7.389%	124	0.5	1	0	7.327%
35	0.25	0.75	0.75	7.415%	80	1	1	0.25	7.388%	125	1	1	0.5	7.299%
36	1	0.5	0	7.414%	81	0.5	0.75	0.25	7.387%					
37	1	0.25	0.5	7.414%	82	0.75	0.75	1	7.386%					
38	0.5	0.5	0	7.413%	83	0.75	0.25	0.25	7.385%					
39	0	0.75	1	7.413%	84	0	0.25	0.25	7.384%					
40	0.75	0.5	1	7.412%	85	0	0	0.25	7.383%					
41	0.25	0.75	0	7.412%	86	1	1	1	7.383%					
42	0	1	0.5	7.411%	87	0.5	0	0.75	7.383%					
43	0.5	0.5	0.25	7.409%	88	1	0.25	0.25	7.383%					
44	0	0	0.75	7.409%	89	1	0	0.75	7.382%					
45	0.25	0.25	0.25	7.408%	90	0	0.25	0.75	7.382%					

Table D.9: F-measure of the FreqRandom model on a monthly basis, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.5	0.25	20.854%	46	1	0.25	0	20.739%	91	1	0	0.75	20.697%
2	0.75	0.5	0.5	20.839%	47	0.25	1	0.75	20.739%	92	0.75	0.25	0.75	20.695%
3	0.5	0	0	20.827%	48	0	1	0.5	20.739%	93	0	0	0.25	20.694%
4	0.5	0.75	0.75	20.822%	49	1	1	1	20.738%	94	0	0.25	0	20.694%
5	0.5	0.75	0.25	20.801%	50	1	0	1	20.738%	95	0.75	0	0.75	20.693%
6	1	0.5	0.5	20.795%	51	1	0.75	0	20.738%	96	0.5	0.5	0.75	20.692%
7	0	0.5	1	20.795%	52	0.5	1	0	20.737%	97	0	0.75	0	20.691%
8	0.25	0.5	0.25	20.792%	53	0.75	0.25	0.25	20.737%	98	0	0	0.5	20.689%
9	0.5	1	0.25	20.792%	54	0.25	0.75	0.5	20.736%	99	0	0.5	0	20.689%
10	0.75	0.75	0.75	20.790%	55	0.5	0.75	0	20.735%	100	0.25	0.5	1	20.689%
11	0	0.5	0.25	20.789%	56	0.75	0	0.25	20.734%	101	0	0.25	0.25	20.683%
12	0	0.25	0.5	20.786%	57	0.5	0.75	1	20.733%	102	0.75	1	0	20.682%
13	0.25	0	1	20.784%	58	0.75	0.5	0.75	20.732%	103	0	0.25	0.75	20.681%
14	0.75	0.25	0.5	20.778%	59	0.5	0.5	0.5	20.732%	104	0.5	0.25	0.5	20.680%
15	1	1	0	20.775%	60	0.5	0.5	1	20.731%	105	1	0.25	0.5	20.679%
16	0.75	0	0.5	20.775%	61	0.75	0.75	0.5	20.730%	106	1	0.5	0.75	20.679%
17	0.5	1	0.5	20.774%	62	1	0.75	0.5	20.727%	107	1	0.75	1	20.677%
18	1	0	0.5	20.769%	63	0.75	0.5	0	20.727%	108	0.25	1	0.5	20.676%
19	0	1	0	20.768%	64	0.25	0.25	1	20.727%	109	0.5	0.5	0	20.671%
20	0.25	0.25	0.25	20.768%	65	0.5	0.25	0.25	20.727%	110	0.75	1	0.25	20.668%
21	0.5	0	0.25	20.767%	66	0.5	0.25	0	20.726%	111	0.75	1	0.5	20.667%
22	1	0.5	0.25	20.765%	67	0.25	0.25	0.75	20.723%	112	1	0.5	1	20.667%
23	0.5	1	1	20.765%	68	0.75	0.25	1	20.722%	113	1	0.75	0.75	20.666%
24	1	1	0.5	20.764%	69	0.75	0.75	1	20.722%	114	0.5	0.5	0.25	20.664%
25	0.25	1	1	20.764%	70	0.25	1	0	20.720%	115	0.5	1	0.75	20.662%
26	0	1	0.25	20.762%	71	0.25	0.5	0.5	20.720%	116	1	0.25	0.25	20.660%
27	0	0.75	1	20.761%	72	0.25	0.75	1	20.718%	117	0.5	0.25	1	20.655%
28	0.75	1	1	20.758%	73	0.25	0.75	0	20.717%	118	0.25	0.25	0.5	20.653%
29	0	0	0	20.757%	74	0	1	0.75	20.713%	119	0.25	0	0.25	20.653%
30	0	0.75	0.25	20.757%	75	0	0.25	1	20.711%	120	0	0.75	0.75	20.641%
31	0	0	1	20.755%	76	0.25	0.25	0	20.710%	121	1	1	0.75	20.638%
32	1	0.5	0	20.755%	77	0.25	0	0.75	20.709%	122	0.75	1	0.75	20.638%
33	0.5	0.75	0.5	20.755%	78	0.25	0	0	20.708%	123	0	0.5	0.5	20.631%
34	0.5	0	1	20.753%	79	0.25	0.5	0.75	20.707%	124	0.75	0	1	20.628%
35	0.5	0	0.5	20.753%	80	0	0	0.75	20.707%	125	0.25	1	0.25	20.622%
36	0.75	0	0	20.751%	81	1	0.25	0.75	20.706%					
37	0	0.5	0.75	20.750%	82	1	1	0.25	20.706%					
38	0.75	0.75	0	20.749%	83	0.25	0.75	0.25	20.705%					
39	1	0.25	1	20.744%	84	0.75	0.5	1	20.705%					
40	0.5	0	0.75	20.744%	85	0.25	0.5	0	20.704%					
41	0	0.75	0.5	20.742%	86	1	0.75	0.25	20.704%					
42	1	0	0.25	20.742%	87	0.25	0	0.5	20.702%					
43	0	1	1	20.740%	88	0.25	0.75	0.75	20.699%					
44	0.75	0.75	0.25	20.740%	89	0.5	0.25	0.75	20.699%					
45	1	0	0	20.739%	90	0.75	0.25	0	20.697%					

Table D.10: F-measure of the FreqLCA model over the whole simulation period, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	0.25	0	19.921%	46	1	0.75	0.25	19.892%	91	0	0.5	0.75	19.814%
2	0	1	0	19.920%	47	0.25	0	0	19.891%	92	0.75	0.75	0.75	19.814%
3	0.75	0	0.5	19.918%	48	0.75	1	0	19.890%	93	0	0.25	0.5	19.809%
4	0	0	0	19.918%	49	1	0.5	0.25	19.887%	94	0	0.75	0.75	19.809%
5	0.75	0	0.75	19.917%	50	0.25	0.25	0	19.886%	95	1	0.5	0.75	19.809%
6	0	0.25	0	19.917%	51	0.5	1	0.5	19.885%	96	0.25	0.25	0.75	19.807%
7	1	0	0	19.916%	52	1	1	0.25	19.885%	97	0	0.25	0.75	19.804%
8	1	0.75	0	19.915%	53	0.5	0.75	0.5	19.884%	98	0.25	0.75	0.75	19.804%
9	0.75	0.25	0	19.915%	54	0.75	0.5	0.25	19.883%	99	0.5	0.25	0.75	19.804%
10	1	0	0.25	19.914%	55	0	1	0.25	19.882%	100	0.25	0.5	0.75	19.802%
11	0.5	0	0.25	19.914%	56	0.25	1	0.25	19.880%	101	0.75	0.5	0.75	19.802%
12	0	0	0.25	19.914%	57	0.5	0.5	0.25	19.878%	102	0.25	1	0.75	19.800%
13	1	0.5	0	19.914%	58	0	0.75	0.25	19.877%	103	0.5	1	0.75	19.800%
14	1	0	0.75	19.912%	59	0	0.5	0.25	19.877%	104	0.75	0.25	0.75	19.796%
15	0.5	0	0.75	19.910%	60	0.25	0.5	0.25	19.875%	105	0.5	1	1	19.788%
16	0.5	0	1	19.909%	61	0.75	0.75	0.25	19.870%	106	1	0.25	1	19.785%
17	0.25	0	0.25	19.909%	62	0.75	0.25	0.25	19.870%	107	0.5	0.5	0.75	19.783%
18	1	0	0.5	19.909%	63	1	0.25	0.25	19.870%	108	0.25	0.25	1	19.783%
19	0.5	0	0	19.908%	64	0.75	1	0.5	19.867%	109	0.75	0.25	1	19.769%
20	0.5	1	0	19.907%	65	0.25	1	0.5	19.866%	110	0	0.25	1	19.766%
21	0.75	0	0	19.906%	66	0.25	0.75	0.25	19.865%	111	0.5	0.25	1	19.765%
22	0.75	0.5	0	19.906%	67	1	0.75	0.5	19.861%	112	0.5	0.75	1	19.765%
23	0.5	0.25	0	19.906%	68	0	0.5	0.5	19.860%	113	0.75	0.75	1	19.764%
24	0	0.5	0	19.905%	69	0.5	0.25	0.5	19.858%	114	0	0.75	1	19.764%
25	0	0	1	19.905%	70	0.5	0.25	0.25	19.857%	115	0.25	1	1	19.757%
26	1	0	1	19.905%	71	0.25	0.25	0.25	19.856%	116	0.25	0.75	1	19.755%
27	0	0	0.75	19.904%	72	0.25	0.75	0.5	19.854%	117	0	0.5	1	19.752%
28	0	0.75	0	19.904%	73	0	0.25	0.25	19.854%	118	0.5	0.5	1	19.749%
29	0.25	0.75	0	19.904%	74	1	1	0.5	19.852%	119	0.75	1	1	19.748%
30	0.25	0	1	19.903%	75	0.25	0.5	0.5	19.852%	120	1	0.75	1	19.744%
31	0.5	0.75	0.25	19.903%	76	0	1	0.5	19.851%	121	0.75	0.5	1	19.735%
32	0.5	0.75	0	19.903%	77	0.5	0.5	0.5	19.848%	122	0.25	0.5	1	19.735%
33	0.5	0	0.5	19.901%	78	0	0.75	0.5	19.848%	123	0	1	1	19.721%
34	0.25	1	0	19.901%	79	1	0.25	0.5	19.846%	124	1	1	1	19.715%
35	0.75	0	0.25	19.900%	80	0.75	0.5	0.5	19.844%	125	1	0.5	1	19.698%
36	0.25	0	0.5	19.900%	81	0.75	0.25	0.5	19.843%					
37	0.75	1	0.25	19.900%	82	1	0.5	0.5	19.841%					
38	1	1	0	19.900%	83	0	1	0.75	19.840%					
39	0.75	0.75	0	19.899%	84	0.5	0.75	0.75	19.837%					
40	0.25	0.5	0	19.898%	85	0.75	0.75	0.5	19.836%					
41	0.75	0	1	19.897%	86	1	0.25	0.75	19.828%					
42	0.5	0.5	0	19.897%	87	0.75	1	0.75	19.823%					
43	0.25	0	0.75	19.897%	88	0.25	0.25	0.5	19.821%					
44	0	0	0.5	19.894%	89	1	0.75	0.75	19.819%					
45	0.5	1	0.25	19.894%	90	1	1	0.75	19.816%					

Table D.11: F-measure of the FreqMax model over the whole simulation period, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0	0.25	20.997%	46	0.25	0.25	1	20.882%	91	0.75	0.25	0.5	20.842%
2	0.75	1	0.5	20.993%	47	0.25	1	0	20.882%	92	0	0.5	0.5	20.841%
3	0.75	1	0.25	20.985%	48	0.75	0.75	0.5	20.882%	93	0.25	0.75	0.25	20.841%
4	0.75	0.5	1	20.972%	49	0	0.25	0	20.880%	94	0.75	0.5	0.25	20.839%
5	0.75	0.25	0.75	20.961%	50	0.25	0.75	1	20.878%	95	1	0.75	0.75	20.839%
6	1	0.25	1	20.953%	51	0	0.5	1	20.878%	96	1	0.5	1	20.839%
7	0.25	0.5	0.75	20.948%	52	1	0	0	20.877%	97	1	0.25	0.75	20.839%
8	0.5	0.75	0.5	20.945%	53	0.5	0.75	0	20.877%	98	0.5	0.25	0.25	20.838%
9	0.25	0.75	0.5	20.941%	54	1	1	0.25	20.876%	99	0.25	0	0.5	20.838%
10	1	0	0.5	20.939%	55	1	0.5	0.25	20.876%	100	1	1	0.5	20.837%
11	0.25	1	0.25	20.936%	56	0.5	0.75	1	20.875%	101	0.25	0.75	0	20.837%
12	1	0.75	0.25	20.935%	57	0.75	0	0.25	20.875%	102	0.5	0.75	0.25	20.837%
13	0.5	0.5	0.25	20.933%	58	0.25	0.25	0.25	20.875%	103	0.25	1	0.75	20.835%
14	0.5	0.5	0.75	20.929%	59	0.75	0	1	20.874%	104	1	0.25	0	20.834%
15	0	0.5	0.25	20.927%	60	0	0	0	20.870%	105	0.5	0.75	0.75	20.834%
16	1	0	1	20.926%	61	0.5	0.25	1	20.870%	106	0.25	0.25	0.75	20.834%
17	0	0.25	0.25	20.921%	62	0.25	0	0.25	20.869%	107	0.75	0.5	0.75	20.823%
18	0.25	0.25	0	20.921%	63	0.75	0.75	0.25	20.867%	108	0.25	0.25	0.5	20.823%
19	0.75	0.75	1	20.921%	64	0.5	0	0.5	20.867%	109	0	0.75	0	20.819%
20	0.25	1	0.5	20.919%	65	0.75	0.25	0.25	20.867%	110	0	0.75	1	20.819%
21	0	1	1	20.919%	66	0.75	1	1	20.866%	111	0	0.75	0.75	20.817%
22	0.75	0	0.5	20.917%	67	0.5	1	0.75	20.865%	112	0	0.25	0.5	20.812%
23	0.25	0.5	0.25	20.915%	68	0	0	0.75	20.864%	113	0.75	0.25	1	20.812%
24	0	1	0.25	20.915%	69	0.5	0.25	0	20.862%	114	0.5	0.5	0	20.810%
25	0.25	0.5	0.5	20.914%	70	0.75	1	0	20.861%	115	0.5	0	0.25	20.808%
26	0	0	1	20.914%	71	0.5	0	0	20.860%	116	0.5	0	1	20.806%
27	0.75	0	0.75	20.914%	72	0.5	1	0.5	20.858%	117	0.5	1	0.25	20.803%
28	1	0.5	0.5	20.909%	73	0	0.75	0.5	20.858%	118	0.75	0.75	0	20.802%
29	0	0.5	0.75	20.907%	74	0	0.5	0	20.858%	119	0	1	0.5	20.791%
30	0	0.25	0.75	20.906%	75	1	0.75	0.5	20.857%	120	0.5	0	0.75	20.783%
31	0.5	1	1	20.904%	76	0.75	0.25	0	20.855%	121	0.25	0	1	20.781%
32	0.25	0	0.75	20.904%	77	0.25	1	1	20.855%	122	0.75	0.5	0	20.768%
33	1	1	1	20.902%	78	0	0.75	0.25	20.854%	123	1	0.5	0	20.765%
34	0.25	0.5	0	20.901%	79	0.75	0	0	20.853%	124	0.5	0.5	1	20.763%
35	0	1	0.75	20.901%	80	1	0.5	0.75	20.852%	125	0.5	0.25	0.5	20.739%
36	0	0.25	1	20.899%	81	0.75	0.5	0.5	20.850%					
37	1	0.25	0.5	20.896%	82	0.75	0.75	0.75	20.850%					
38	1	0.75	1	20.895%	83	0.25	0.75	0.75	20.850%					
39	0	1	0	20.891%	84	1	1	0	20.849%					
40	0	0	0.5	20.890%	85	1	0	0.75	20.848%					
41	0.5	0.5	0.5	20.889%	86	1	0.25	0.25	20.847%					
42	0.5	0.25	0.75	20.885%	87	0.75	1	0.75	20.844%					
43	1	1	0.75	20.885%	88	1	0	0.25	20.844%					
44	1	0.75	0	20.885%	89	0.25	0	0	20.843%					
45	0.5	1	0	20.883%	90	0.25	0.5	1	20.842%					

Table D.12: F-measure of the FreqLogit model over the whole simulation period, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.5	0.25	20.990%	46	0	0	0.5	20.898%	91	0	0	0.25	20.862%
2	0.75	0.5	0.75	20.985%	47	1	0.5	0.25	20.896%	92	0.5	0.25	0.25	20.862%
3	0.5	0.25	0	20.982%	48	0	0.5	0.25	20.895%	93	0.5	0	0.25	20.860%
4	0.25	0	0.5	20.961%	49	0	0.75	0	20.895%	94	0.5	0.75	0.75	20.860%
5	0.75	0.75	0.75	20.957%	50	0.75	0.25	0.25	20.893%	95	1	1	0.25	20.859%
6	0.25	0	1	20.950%	51	0.75	0.5	0	20.893%	96	0.5	0.75	0.25	20.853%
7	0.75	0.75	0.5	20.949%	52	1	0	0.25	20.890%	97	0.5	0.5	1	20.853%
8	0.25	0	0.75	20.941%	53	0.25	0.5	0.5	20.888%	98	0.25	0.25	0.75	20.853%
9	1	0.5	0	20.941%	54	0.25	0.25	1	20.888%	99	0.25	0.5	0	20.851%
10	0.5	0.25	0.5	20.941%	55	0	0.75	1	20.887%	100	1	0.25	0.75	20.850%
11	1	0	0.5	20.940%	56	0.75	0	0	20.887%	101	0	1	0	20.850%
12	1	0.5	0.5	20.936%	57	0	0.25	0.75	20.887%	102	1	0.75	0.5	20.850%
13	0	0.75	0.5	20.935%	58	0.25	1	1	20.885%	103	0.75	1	0.25	20.849%
14	1	0.5	1	20.934%	59	0.5	0	0.75	20.885%	104	0	1	0.5	20.841%
15	0.25	1	0.75	20.930%	60	1	1	0.75	20.883%	105	0.5	0.75	1	20.840%
16	0.25	0	0.25	20.928%	61	0.5	0.25	1	20.883%	106	1	0	0	20.840%
17	0.75	0	0.5	20.926%	62	1	0.5	0.75	20.883%	107	1	1	0	20.839%
18	0.25	0.75	1	20.925%	63	0	0.5	0.5	20.883%	108	0	0.25	0.5	20.835%
19	0	0.25	1	20.925%	64	0.75	0.25	0	20.882%	109	1	1	0.5	20.834%
20	0.25	0.75	0.5	20.924%	65	0	0.25	0.25	20.881%	110	0.25	0.5	1	20.833%
21	0.5	0	0	20.923%	66	0	0.5	1	20.880%	111	0.25	0.75	0.25	20.833%
22	0.25	0.75	0	20.922%	67	0.75	0.5	1	20.879%	112	0.5	0	1	20.832%
23	0.75	0.5	0.5	20.922%	68	1	0.75	0	20.879%	113	0.25	0.25	0	20.831%
24	1	0.25	0.5	20.921%	69	0	0	0.75	20.878%	114	0.5	0	0.5	20.830%
25	0.5	0.5	0	20.921%	70	1	0	1	20.878%	115	0.75	0	0.25	20.829%
26	0.75	1	0	20.920%	71	0	1	0.75	20.878%	116	1	0.75	1	20.823%
27	0.75	0.75	0.25	20.919%	72	0.5	0.5	0.5	20.876%	117	0	1	0.25	20.822%
28	0.5	0.75	0.5	20.918%	73	0.5	0.25	0.75	20.876%	118	0.75	0	1	20.821%
29	0	0	0	20.913%	74	0.5	0.75	0	20.874%	119	0	0.75	0.25	20.815%
30	0	0.25	0	20.913%	75	1	0.25	0.25	20.874%	120	0.5	1	0	20.813%
31	0.75	1	0.75	20.912%	76	0.25	0.25	0.5	20.874%	121	1	0.25	1	20.808%
32	1	0	0.75	20.911%	77	0.75	0	0.75	20.872%	122	0	0.5	0.75	20.808%
33	0.25	0.75	0.75	20.910%	78	0.75	0.25	0.5	20.872%	123	1	1	1	20.807%
34	0.75	0.25	0.75	20.907%	79	0.75	1	0.5	20.871%	124	0.5	1	0.5	20.788%
35	1	0.75	0.25	20.906%	80	0.5	0.5	0.25	20.871%	125	0	1	1	20.764%
36	1	0.75	0.75	20.904%	81	0.25	1	0	20.870%					
37	0.75	1	1	20.904%	82	0.25	0	0	20.870%					
38	0.75	0.25	1	20.904%	83	0.5	1	0.75	20.870%					
39	0.25	1	0.5	20.903%	84	1	0.25	0	20.868%					
40	0.5	0.5	0.75	20.902%	85	0	0.75	0.75	20.868%					
41	0.25	0.25	0.25	20.901%	86	0.75	0.75	0	20.866%					
42	0.25	0.5	0.25	20.900%	87	0	0	1	20.866%					
43	0.5	1	0.25	20.900%	88	0.75	0.75	1	20.865%					
44	0	0.5	0	20.899%	89	0.25	0.5	0.75	20.865%					
45	0.25	1	0.25	20.898%	90	0.5	1	1	20.864%					

Table D.13: F-measure of the FreqRandom model over the whole simulation period, for the all-previous topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.5	0.75	0.508%	46	0	0	1	0.491%	91	0.5	0.75	0.75	0.480%
2	0.75	0.75	0.75	0.508%	47	0	0.5	0	0.491%	92	0.25	0.5	1	0.480%
3	0	0	0.75	0.508%	48	0.25	0.75	0.5	0.491%	93	0.75	0.5	0.25	0.479%
4	0	1	0.25	0.508%	49	0.75	0.25	0	0.490%	94	0	0	0	0.479%
5	1	0.75	1	0.505%	50	0.25	0.25	0.75	0.490%	95	0.25	0.75	0	0.479%
6	0.75	0	1	0.505%	51	1	0.25	0.25	0.490%	96	0.75	0	0.75	0.479%
7	1	0.75	0.75	0.505%	52	0	0.75	0.5	0.490%	97	0.5	1	0.75	0.479%
8	1	0	0.5	0.504%	53	0.75	0.25	1	0.489%	98	0.5	0.5	0.25	0.478%
9	0.25	1	1	0.504%	54	0	1	0	0.489%	99	0	0.25	1	0.478%
10	0.25	1	0	0.502%	55	0.25	1	0.25	0.489%	100	0.25	0.5	0.25	0.477%
11	0.75	0	0.25	0.502%	56	0.5	0.75	0	0.488%	101	0.75	0.25	0.75	0.477%
12	0.5	0.5	0.75	0.502%	57	0	0.5	1	0.488%	102	0.75	1	0.25	0.477%
13	0.25	0	0.25	0.501%	58	0	0.75	0.25	0.488%	103	0.75	1	0.5	0.477%
14	0.5	0.25	0	0.501%	59	1	0.75	0.25	0.488%	104	0.75	0.5	0.5	0.477%
15	0.75	1	0	0.501%	60	0.25	1	0.75	0.488%	105	0.5	0	0.75	0.476%
16	0	0.75	1	0.501%	61	0.5	0	1	0.488%	106	1	0	0	0.474%
17	0.5	0	0.25	0.500%	62	0.25	0.75	1	0.488%	107	0.75	0.25	0.25	0.474%
18	0.25	0.25	0.25	0.500%	63	1	0.75	0	0.487%	108	1	0.25	0	0.474%
19	0.75	0	0.5	0.499%	64	0	0.75	0	0.487%	109	0.5	0.5	1	0.473%
20	0.75	0.25	0.5	0.499%	65	1	1	0.75	0.487%	110	0.25	0.75	0.25	0.473%
21	0.75	0	0	0.498%	66	0.25	1	0.5	0.486%	111	1	0	0.25	0.473%
22	0.5	0	0.5	0.498%	67	1	0.25	1	0.486%	112	1	0.5	0	0.472%
23	0.5	1	0	0.498%	68	0.75	0.75	1	0.486%	113	1	0	1	0.472%
24	0.25	0	0	0.498%	69	0.5	0.25	1	0.486%	114	0.75	0.5	0	0.472%
25	1	0.5	0.5	0.497%	70	0.5	0.25	0.5	0.486%	115	1	0.25	0.5	0.471%
26	0	0.25	0.5	0.497%	71	1	0.75	0.5	0.485%	116	0	0.25	0.75	0.471%
27	0.25	0.75	0.75	0.495%	72	0.25	0.25	1	0.485%	117	0	0.25	0.25	0.471%
28	0	1	1	0.495%	73	1	0.5	1	0.485%	118	0.75	0.5	0.75	0.471%
29	0.25	0.25	0	0.495%	74	0	1	0.5	0.485%	119	0	0.75	0.75	0.470%
30	0	0	0.25	0.495%	75	1	1	0.5	0.485%	120	0.25	0	0.5	0.470%
31	0.25	0.5	0.75	0.494%	76	0.75	0.75	0.25	0.485%	121	0.5	0.75	1	0.469%
32	0.75	1	1	0.494%	77	0.75	0.75	0.5	0.485%	122	0	0.5	0.25	0.469%
33	0	1	0.75	0.494%	78	0.5	1	1	0.485%	123	1	1	1	0.468%
34	0.25	0	1	0.494%	79	1	1	0.25	0.484%	124	0	0.5	0.5	0.464%
35	0.5	0.5	0	0.493%	80	1	0.5	0.75	0.484%	125	0.25	0	0.75	0.459%
36	1	0.25	0.75	0.493%	81	0.5	0	0	0.484%					
37	0.75	0.75	0	0.493%	82	1	0	0.75	0.484%					
38	0.5	0.75	0.5	0.493%	83	1	0.5	0.25	0.483%					
39	0.5	0.25	0.75	0.492%	84	1	1	0	0.483%					
40	0.5	1	0.25	0.492%	85	0.25	0.25	0.5	0.483%					
41	0.25	0.5	0.5	0.492%	86	0	0.25	0	0.482%					
42	0	0	0.5	0.492%	87	0.5	0.25	0.25	0.482%					
43	0.5	1	0.5	0.492%	88	0.75	1	0.75	0.481%					
44	0.5	0.75	0.25	0.491%	89	0.5	0.5	0.5	0.480%					
45	0.75	0.5	1	0.491%	90	0.25	0.5	0	0.480%					

Table D.14: F-measure of the FreqLCA model on a weekly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
-	0	0	0	NaN	-	0.25	0	1	NaN	-	0.75	0	0.75	NaN
-	0	0.25	0	NaN	-	0.25	0.25	1	NaN	-	0.75	0.25	0.75	NaN
-	0	0.5	0	NaN	-	0.25	0.5	1	NaN	-	0.75	0.5	0.75	NaN
-	0	0.75	0	NaN	-	0.25	0.75	1	NaN	-	0.75	0.75	0.75	NaN
-	0	1	0	NaN	-	0.25	1	1	NaN	-	0.75	1	0.75	NaN
-	0	0	0.25	NaN	-	0.5	0	0	NaN	-	0.75	0	1	NaN
-	0	0.25	0.25	NaN	-	0.5	0.25	0	NaN	-	0.75	0.25	1	NaN
-	0	0.5	0.25	NaN	-	0.5	0.5	0	NaN	-	0.75	0.5	1	NaN
-	0	0.75	0.25	NaN	-	0.5	0.75	0	NaN	-	0.75	0.75	1	NaN
-	0	1	0.25	NaN	-	0.5	1	0	NaN	-	0.75	1	1	NaN
-	0	0	0.5	NaN	-	0.5	0	0.25	NaN	-	1	0	0	NaN
-	0	0.25	0.5	NaN	-	0.5	0.25	0.25	NaN	-	1	0.25	0	NaN
-	0	0.5	0.5	NaN	-	0.5	0.5	0.25	NaN	-	1	0.5	0	NaN
-	0	0.75	0.5	NaN	-	0.5	0.75	0.25	NaN	-	1	0.75	0	NaN
-	0	1	0.5	NaN	-	0.5	1	0.25	NaN	-	1	1	0	NaN
-	0	0	0.75	NaN	-	0.5	0	0.5	NaN	-	1	0	0.25	NaN
-	0	0.25	0.75	NaN	-	0.5	0.25	0.5	NaN	-	1	0.25	0.25	NaN
-	0	0.5	0.75	NaN	-	0.5	0.5	0.5	NaN	-	1	0.5	0.25	NaN
-	0	0.75	0.75	NaN	-	0.5	0.75	0.5	NaN	-	1	0.75	0.25	NaN
-	0	1	0.75	NaN	-	0.5	1	0.5	NaN	-	1	1	0.25	NaN
-	0	0	1	NaN	-	0.5	0	0.75	NaN	-	1	0	0.5	NaN
-	0	0.25	1	NaN	-	0.5	0.25	0.75	NaN	-	1	0.25	0.5	NaN
-	0	0.5	1	NaN	-	0.5	0.5	0.75	NaN	-	1	0.5	0.5	NaN
-	0	0.75	1	NaN	-	0.5	0.75	0.75	NaN	-	1	0.75	0.5	NaN
-	0	1	1	NaN	-	0.5	1	0.75	NaN	-	1	1	0.5	NaN
-	0.25	0	0	NaN	-	0.5	0	1	NaN	-	1	0	0.75	NaN
-	0.25	0.25	0	NaN	-	0.5	0.25	1	NaN	-	1	0.25	0.75	NaN
-	0.25	0.5	0	NaN	-	0.5	0.5	1	NaN	-	1	0.5	0.75	NaN
-	0.25	0.75	0	NaN	-	0.5	0.75	1	NaN	-	1	0.75	0.75	NaN
-	0.25	1	0	NaN	-	0.5	1	1	NaN	-	1	1	0.75	NaN
-	0.25	0	0.25	NaN	-	0.75	0	0	NaN	-	1	0	1	NaN
-	0.25	0.25	0.25	NaN	-	0.75	0.25	0	NaN	-	1	0.25	1	NaN
-	0.25	0.5	0.25	NaN	-	0.75	0.5	0	NaN	-	1	0.5	1	NaN
-	0.25	0.75	0.25	NaN	-	0.75	0.75	0	NaN	-	1	0.75	1	NaN
-	0.25	1	0.25	NaN	-	0.75	1	0	NaN	-	1	1	1	NaN
-	0.25	0	0.5	NaN	-	0.75	0	0.25	NaN					
-	0.25	0.25	0.5	NaN	-	0.75	0.25	0.25	NaN					
-	0.25	0.5	0.5	NaN	-	0.75	0.5	0.25	NaN					
-	0.25	0.75	0.5	NaN	-	0.75	0.75	0.25	NaN					
-	0.25	1	0.5	NaN	-	0.75	1	0.25	NaN					
-	0.25	0	0.75	NaN	-	0.75	0	0.5	NaN					
-	0.25	0.25	0.75	NaN	-	0.75	0.25	0.5	NaN					
-	0.25	0.5	0.75	NaN	-	0.75	0.5	0.5	NaN					
-	0.25	0.75	0.75	NaN	-	0.75	0.75	0.5	NaN					
-	0.25	1	0.75	NaN	-	0.75	1	0.5	NaN					

Table D.15: F-measure of the FreqMax model on a weekly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	1	0.5	0.515%	46	0.25	0.75	0.75	0.496%	91	0.25	0.25	0.5	0.484%
2	0.25	0.5	0.25	0.515%	47	0	1	1	0.496%	92	1	0.75	0.75	0.484%
3	0.75	0	0.75	0.515%	48	0.75	1	0.75	0.495%	93	0	0.25	0.5	0.483%
4	1	0.25	0	0.513%	49	0.25	0.75	0	0.495%	94	0	0.25	0.25	0.483%
5	1	0.75	1	0.513%	50	0	1	0.5	0.495%	95	0.75	0.5	0	0.483%
6	0.5	0.75	0	0.513%	51	0.75	0.25	1	0.494%	96	0.5	0.75	1	0.482%
7	0.25	1	0.25	0.512%	52	0	0.25	0.75	0.494%	97	0.75	0.75	1	0.482%
8	0.5	1	0	0.509%	53	0	0	0.5	0.494%	98	0	1	0.25	0.482%
9	1	0.75	0.25	0.509%	54	0.5	0.5	0	0.494%	99	0.5	0.5	1	0.481%
10	0.75	0.5	0.25	0.508%	55	0.25	0	0	0.494%	100	0	0	0	0.480%
11	1	0.5	0.25	0.508%	56	0.5	0.5	0.5	0.494%	101	1	0.75	0.5	0.480%
12	0.25	0.75	0.25	0.507%	57	1	0.5	1	0.493%	102	1	0.5	0.75	0.480%
13	0.75	0	0.5	0.507%	58	0.5	0	0	0.493%	103	1	0.75	0	0.480%
14	0.5	0.25	0.75	0.506%	59	0.75	1	0.25	0.493%	104	0.25	0.5	0	0.479%
15	1	0	0	0.506%	60	1	1	1	0.492%	105	1	0	0.75	0.479%
16	0.25	1	0.75	0.505%	61	0	0.75	0.5	0.492%	106	0.25	1	1	0.479%
17	0	0	0.25	0.505%	62	0.5	0.5	0.75	0.492%	107	0	0.25	0	0.477%
18	0.25	0.75	0.5	0.504%	63	1	1	0.75	0.492%	108	0	0	1	0.477%
19	0.25	0.5	0.5	0.504%	64	0.25	0.75	1	0.492%	109	0.75	1	1	0.476%
20	0.5	1	1	0.504%	65	0.5	0	0.5	0.491%	110	0.5	0.25	0	0.476%
21	0.5	1	0.25	0.504%	66	0	0.75	0	0.490%	111	0.75	0	0	0.476%
22	0.5	1	0.5	0.503%	67	1	0	0.25	0.490%	112	0	1	0.75	0.475%
23	0	1	0	0.502%	68	1	1	0	0.490%	113	0.75	0.75	0.75	0.475%
24	1	0.25	1	0.502%	69	1	1	0.5	0.490%	114	0.75	0.75	0.25	0.474%
25	1	0.5	0.5	0.502%	70	1	0.25	0.75	0.490%	115	1	1	0.25	0.473%
26	0.75	0.75	0.5	0.501%	71	1	0.5	0	0.489%	116	0	0.5	0	0.473%
27	0.5	0.25	1	0.501%	72	0	0.75	0.25	0.489%	117	0	0.75	1	0.472%
28	0.5	0	1	0.500%	73	0.5	1	0.75	0.489%	118	0.25	0.25	0.75	0.472%
29	0.25	0	0.75	0.500%	74	0.5	0.5	0.25	0.488%	119	0.25	0	0.5	0.470%
30	0	0	0.75	0.500%	75	0	0.25	1	0.488%	120	0.75	1	0	0.469%
31	0.75	0.25	0.5	0.500%	76	0.25	0.25	0	0.488%	121	0.75	0.75	0	0.469%
32	0.5	0.75	0.5	0.500%	77	0.25	0.5	0.75	0.487%	122	0.75	0.5	0.5	0.466%
33	0.5	0.25	0.25	0.500%	78	0.25	0.25	0.25	0.487%	123	0.5	0	0.25	0.464%
34	0	0.5	1	0.500%	79	1	0.25	0.5	0.487%	124	0.25	0	1	0.462%
35	0.75	0.25	0.75	0.500%	80	0	0.75	0.75	0.487%	125	0.75	0.25	0	NaN
36	1	0	1	0.499%	81	0.75	0	0.25	0.487%					
37	0.5	0.25	0.5	0.498%	82	0.25	0	0.25	0.487%					
38	0.75	0.5	1	0.498%	83	1	0.25	0.25	0.486%					
39	0.75	0	1	0.498%	84	0.75	0.25	0.25	0.486%					
40	1	0	0.5	0.498%	85	0	0.5	0.25	0.485%					
41	0.25	1	0.5	0.497%	86	0.25	0.5	1	0.485%					
42	0.25	1	0	0.497%	87	0.5	0.75	0.75	0.485%					
43	0.5	0	0.75	0.497%	88	0	0.5	0.5	0.485%					
44	0.5	0.75	0.25	0.497%	89	0.25	0.25	1	0.485%					
45	0.75	0.5	0.75	0.496%	90	0	0.5	0.75	0.485%					

Table D.16: F-measure of the FreqLogit model on a weekly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	0	0	0.526%	46	0.25	0.5	0.25	0.498%	91	0.5	0.25	0.5	0.486%
2	0.5	0.25	0	0.521%	47	0	0.75	0.5	0.497%	92	0.5	0.5	0.5	0.486%
3	0.25	0.5	0.5	0.520%	48	1	0	1	0.497%	93	0	0.75	0	0.485%
4	0.75	0.5	1	0.519%	49	0	0.5	0.25	0.497%	94	0.25	0.75	0.75	0.485%
5	0	0.25	0	0.518%	50	1	0.25	0.5	0.497%	95	0	0.25	0.25	0.485%
6	0.75	0.25	0.75	0.515%	51	0.75	0	0.25	0.497%	96	0.25	0	1	0.485%
7	0.75	1	1	0.515%	52	0.25	0.25	0	0.497%	97	0	1	0.25	0.485%
8	0.5	1	0.75	0.514%	53	0	1	0.75	0.496%	98	1	0.75	0.25	0.484%
9	0	0	0.75	0.513%	54	1	0.75	0.75	0.496%	99	0.25	1	0.75	0.484%
10	1	0.25	1	0.512%	55	0	0.5	0.5	0.496%	100	0.75	0.25	0.25	0.484%
11	0.75	1	0.25	0.511%	56	0.25	0.5	0.75	0.495%	101	0.25	0.25	0.75	0.483%
12	0.5	0.5	0.25	0.510%	57	0.75	0.5	0	0.495%	102	1	1	1	0.483%
13	1	0.5	0	0.509%	58	0	0.5	0	0.495%	103	0.75	0.5	0.25	0.482%
14	0.75	0	0.5	0.509%	59	0.75	0.75	0.75	0.495%	104	0.25	0	0.5	0.482%
15	0	1	0.5	0.507%	60	0.25	0	0	0.495%	105	0.75	0.5	0.5	0.480%
16	0.75	0	1	0.506%	61	0.75	0.75	0.5	0.495%	106	0.25	0.75	0.5	0.480%
17	0.75	1	0.5	0.506%	62	1	0.25	0.25	0.495%	107	0.5	0.75	0.25	0.480%
18	0.5	0.75	1	0.505%	63	0.25	0.25	0.5	0.494%	108	0	0.5	1	0.480%
19	1	1	0.5	0.505%	64	0	0.75	0.75	0.494%	109	1	0	0.25	0.480%
20	0	0.25	0.5	0.504%	65	0.75	1	0.75	0.493%	110	1	0	0.75	0.480%
21	0.75	0.25	1	0.504%	66	0	0	0.5	0.492%	111	0.75	0.75	0	0.479%
22	0.5	0.25	1	0.504%	67	0.75	0.5	0.75	0.492%	112	0.75	0	0.75	0.479%
23	0.25	1	0.25	0.503%	68	0.5	0.75	0.5	0.491%	113	0.5	1	0	0.478%
24	0.5	0	0.75	0.503%	69	0.5	1	1	0.491%	114	0.25	0.75	0	0.477%
25	1	1	0.25	0.502%	70	0.25	0.75	1	0.491%	115	0.5	0.5	1	0.476%
26	1	0.25	0.75	0.502%	71	0.5	0.75	0	0.491%	116	1	0.5	0.75	0.476%
27	0.5	0	1	0.502%	72	1	0.75	1	0.491%	117	0.25	0.5	0	0.476%
28	1	0	0.5	0.502%	73	0	0	1	0.491%	118	0	0.25	0.75	0.475%
29	0.5	0.5	0	0.501%	74	1	0.75	0.5	0.490%	119	0.5	0	0.5	0.474%
30	0.5	0	0	0.501%	75	0.25	1	0.5	0.490%	120	1	1	0.75	0.474%
31	0	1	0	0.501%	76	1	0.5	1	0.489%	121	0.5	0.25	0.75	0.473%
32	1	0.25	0	0.501%	77	0	0.25	1	0.489%	122	0.5	1	0.25	0.470%
33	0.5	0	0.25	0.500%	78	0.25	0.5	1	0.489%	123	0	0	0	0.466%
34	1	0.75	0	0.500%	79	0.25	0.75	0.25	0.489%	124	0.5	0.25	0.25	0.461%
35	0.75	0.25	0.5	0.500%	80	0.25	1	0	0.488%	125	0.5	0.5	0.75	NaN
36	0	0.75	1	0.499%	81	0.75	0	0	0.488%					
37	1	1	0	0.499%	82	0.25	0.25	0.25	0.488%					
38	0	1	1	0.499%	83	0.5	1	0.5	0.488%					
39	0.75	0.75	1	0.499%	84	0.25	0	0.25	0.488%					
40	0	0.5	0.75	0.499%	85	1	0.5	0.5	0.487%					
41	1	0.5	0.25	0.498%	86	0.75	1	0	0.487%					
42	0.75	0.25	0	0.498%	87	0.25	0	0.75	0.486%					
43	0	0	0.25	0.498%	88	0.5	0.75	0.75	0.486%					
44	0.25	1	1	0.498%	89	0.75	0.75	0.25	0.486%					
45	0	0.75	0.25	0.498%	90	0.25	0.25	1	0.486%					

Table D.17: F-measure of the FreqRandom model on a weekly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.5	0.75	1.764%	46	0.75	0.25	0	1.698%	91	0	0.25	0.25	1.676%
2	0.5	0	0.25	1.751%	47	0.25	0.25	0	1.698%	92	0.5	0.5	1	1.676%
3	0.75	0	0	1.751%	48	0.25	0	1	1.698%	93	0.5	1	0.5	1.676%
4	0.5	0	0	1.747%	49	0	0	0	1.698%	94	1	0	1	1.674%
5	0	0	0.25	1.741%	50	1	0.5	0.25	1.697%	95	0.75	0.75	0.5	1.674%
6	0.25	0.5	0.5	1.739%	51	1	1	0.25	1.697%	96	0.5	0.5	0.25	1.674%
7	0	1	0.25	1.735%	52	0	0	0.75	1.697%	97	0	0.5	0.25	1.673%
8	1	0.75	1	1.732%	53	0.25	1	0.75	1.696%	98	0	0.75	0.75	1.673%
9	0.5	0.25	0.5	1.731%	54	0.5	0.5	0	1.696%	99	1	0.5	1	1.673%
10	0.75	0.25	0.5	1.729%	55	0.75	0.25	0.25	1.696%	100	0.25	0.75	0.75	1.671%
11	0	0.5	0	1.726%	56	1	0.75	0	1.695%	101	0.5	1	0	1.671%
12	1	0.75	0.75	1.725%	57	0	0	1	1.695%	102	0.75	1	0.25	1.670%
13	0.25	1	1	1.725%	58	1	0.5	0.5	1.694%	103	0.5	0.25	0	1.669%
14	0	1	0.75	1.723%	59	1	0.25	0.25	1.693%	104	0	0.75	0.5	1.669%
15	1	0	0.75	1.722%	60	0.5	0.25	0.75	1.693%	105	0.75	0	0.5	1.668%
16	0.5	0.75	0.75	1.720%	61	0	0.25	1	1.693%	106	0.75	0.25	1	1.667%
17	0.75	0.75	0.75	1.719%	62	0.75	0.75	1	1.692%	107	0	0.25	0	1.667%
18	0.75	0.75	0.25	1.716%	63	0.25	1	0.5	1.692%	108	0.5	0.75	1	1.666%
19	0.5	1	0.25	1.715%	64	0.25	0.25	1	1.691%	109	0.5	1	0.75	1.665%
20	0.25	0.75	0	1.715%	65	0.5	0.25	0.25	1.691%	110	0	0.5	0.5	1.665%
21	0.25	0.75	0.5	1.714%	66	0	0.5	1	1.690%	111	1	1	0.75	1.665%
22	0.75	0	0.75	1.713%	67	0	0.75	0.25	1.689%	112	0.5	0	0.75	1.664%
23	0	1	1	1.712%	68	0.5	0.75	0	1.688%	113	0.5	0.5	0.5	1.663%
24	0.5	0	1	1.712%	69	0	0.25	0.75	1.688%	114	1	0.5	0.75	1.661%
25	0.75	1	0	1.712%	70	0.25	0.5	0.25	1.687%	115	0.75	0.5	0.75	1.660%
26	0	0.75	0	1.712%	71	1	0	0.25	1.687%	116	0.25	0.5	1	1.659%
27	0	1	0	1.711%	72	0.25	0	0.5	1.687%	117	0.75	1	0.75	1.658%
28	0.25	0.25	0.75	1.710%	73	1	0	0	1.687%	118	0.75	0.5	0	1.652%
29	0.75	0	0.25	1.710%	74	1	0.75	0.25	1.687%	119	1	0.25	0.5	1.652%
30	0.5	0	0.5	1.710%	75	0.25	0.75	1	1.685%	120	0.5	1	1	1.651%
31	0.25	0.75	0.25	1.709%	76	0.5	0.25	1	1.685%	121	1	1	0.5	1.651%
32	0.25	0	0.25	1.708%	77	1	0	0.5	1.684%	122	1	0.25	0.75	1.649%
33	0.25	1	0	1.707%	78	0	1	0.5	1.684%	123	0.25	1	0.25	1.648%
34	0	0.25	0.5	1.707%	79	1	1	0	1.683%	124	0.75	1	0.5	1.646%
35	0.75	0	1	1.706%	80	0	0	0.5	1.683%	125	0.25	0.25	0.5	1.630%
36	0.75	1	1	1.705%	81	0.25	0.25	0.25	1.683%					
37	0.75	0.5	0.5	1.704%	82	0.75	0.5	0.25	1.683%					
38	0.25	0.5	0	1.703%	83	0.75	0.5	1	1.682%					
39	1	1	1	1.703%	84	0.25	0	0.75	1.682%					
40	0.5	0.75	0.5	1.703%	85	1	0.5	0	1.681%					
41	0	0.75	1	1.702%	86	1	0.25	0	1.679%					
42	0.75	0.25	0.75	1.701%	87	0.5	0.75	0.25	1.678%					
43	1	0.75	0.5	1.701%	88	0.75	0.75	0	1.678%					
44	0.5	0.5	0.75	1.700%	89	1	0.25	1	1.676%					
45	0.25	0	0	1.699%	90	0.25	0.5	0.75	1.676%					

Table D.18: F-measure of the FreqLCA model on a monthly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.5	0.5	0.25	1.694%	46	0	0	0	1.664%	91	0	0.75	0.25	1.650%
2	0	0.25	1	1.691%	47	0.25	0.5	1	1.664%	92	0.75	0.5	0	1.650%
3	0	0.25	0.75	1.687%	48	1	1	0.25	1.663%	93	0	0	0.75	1.649%
4	1	1	0.5	1.687%	49	1	0	0.75	1.663%	94	0.5	0.25	0	1.649%
5	0.75	0.5	0.25	1.686%	50	1	0	0	1.663%	95	0	1	0	1.648%
6	0	0.75	0	1.686%	51	1	0.5	0.25	1.663%	96	0.75	0.25	0.25	1.648%
7	0.25	1	1	1.686%	52	0.5	0.75	0.25	1.662%	97	0.75	0.75	0	1.648%
8	0	0.75	0.75	1.683%	53	1	0	0.5	1.662%	98	0.75	0.75	0.75	1.648%
9	0.75	0.25	0.75	1.683%	54	0.5	1	0	1.662%	99	0.5	0.75	0.5	1.647%
10	0.5	0	0.25	1.682%	55	0.25	0.75	0	1.662%	100	0.75	0.5	0.5	1.646%
11	0.5	0.25	0.5	1.681%	56	0.25	0.5	0.25	1.661%	101	0	0.25	0.25	1.646%
12	0.75	0.75	0.25	1.681%	57	0.5	0.5	0.5	1.661%	102	0.5	0	0.5	1.646%
13	0.5	0.25	1	1.681%	58	0.75	0.25	0	1.661%	103	0.25	0	0.25	1.645%
14	0.75	0	0.25	1.678%	59	0.25	0.5	0.75	1.660%	104	0.5	0	0	1.645%
15	0	0.5	0.75	1.678%	60	0.5	1	0.75	1.660%	105	1	0.5	0.5	1.644%
16	0	0.5	0.5	1.677%	61	1	1	1	1.659%	106	0.25	0.75	0.5	1.644%
17	0.25	0.25	1	1.677%	62	0.75	0	0.5	1.659%	107	0.75	0.75	1	1.644%
18	0.75	1	0.25	1.676%	63	0.75	1	1	1.659%	108	1	0.25	0	1.643%
19	0.75	1	0.75	1.675%	64	0.5	1	0.5	1.658%	109	0.75	0.75	0.5	1.642%
20	0.5	0.25	0.75	1.675%	65	0.5	0.5	1	1.658%	110	0	0.25	0.5	1.642%
21	0.5	0.25	0.25	1.675%	66	1	0.75	0.5	1.658%	111	1	0	0.25	1.642%
22	0.75	0	0.75	1.674%	67	0	0.25	0	1.658%	112	0.25	0.75	1	1.642%
23	0.75	0.5	0.75	1.674%	68	0	0	1	1.657%	113	0.25	1	0	1.641%
24	0.25	0	0.75	1.672%	69	1	0	1	1.657%	114	1	0.75	1	1.641%
25	0.25	0.25	0.75	1.672%	70	1	0.5	0.75	1.656%	115	1	0.75	0.75	1.639%
26	0.75	0	0	1.672%	71	0	0.5	0	1.656%	116	0.5	0.75	1	1.639%
27	0.25	1	0.25	1.672%	72	0.25	0.25	0.5	1.656%	117	1	0.5	1	1.638%
28	0.75	0.25	0.5	1.672%	73	0	0	0.25	1.656%	118	0.75	0.25	1	1.637%
29	0.5	0.5	0.75	1.671%	74	0.25	1	0.75	1.655%	119	1	0.75	0	1.634%
30	0.25	0.75	0.75	1.670%	75	0.5	0	1	1.655%	120	0.25	0.5	0	1.633%
31	0	1	1	1.669%	76	1	0.25	1	1.655%	121	1	1	0	1.631%
32	1	0.75	0.25	1.669%	77	0.75	0.5	1	1.655%	122	1	0.25	0.5	1.630%
33	0.25	1	0.5	1.669%	78	0.25	0.5	0.5	1.655%	123	0.75	1	0	1.629%
34	0.5	0.75	0	1.669%	79	0.25	0.25	0	1.654%	124	0	0.75	1	1.620%
35	0.5	0.75	0.75	1.668%	80	1	0.25	0.75	1.654%	125	1	0.5	0	1.619%
36	0.25	0	0.5	1.668%	81	0.5	1	0.25	1.653%					
37	0.25	0.25	0.25	1.668%	82	0.25	0	1	1.653%					
38	0	1	0.75	1.667%	83	0	0	0.5	1.653%					
39	0.25	0	0	1.667%	84	0.5	0	0.75	1.653%					
40	0	0.5	0.25	1.667%	85	0	0.75	0.5	1.652%					
41	1	0.25	0.25	1.666%	86	0.5	1	1	1.652%					
42	0	1	0.5	1.666%	87	0.5	0.5	0	1.651%					
43	0.75	0	1	1.665%	88	0	0.5	1	1.651%					
44	0.75	1	0.5	1.665%	89	1	1	0.75	1.651%					
45	0	1	0.25	1.665%	90	0.25	0.75	0.25	1.651%					

Table D.19: F-measure of the FreqMax model on a monthly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.25	1	0.25	1.762%	46	0.25	0.25	1	1.714%	91	1	0.5	1	1.690%
2	0.25	0.75	0.5	1.755%	47	0.75	0.5	1	1.713%	92	0.5	0.75	0.25	1.690%
3	1	0.75	1	1.753%	48	0.25	1	0	1.713%	93	0	0	0	1.690%
4	1	0	0	1.750%	49	0	0.25	1	1.713%	94	0.5	0.25	0	1.689%
5	0.5	0.5	0.5	1.744%	50	0.75	0	0	1.711%	95	1	0.25	0.75	1.688%
6	0.5	1	0.25	1.742%	51	0.5	0.25	0.25	1.711%	96	0.5	0.5	1	1.687%
7	0	0.25	0.25	1.742%	52	0	0.5	1	1.710%	97	0	0.5	0.5	1.686%
8	0.25	0	0.25	1.740%	53	0.75	0.75	0	1.709%	98	0.75	1	1	1.686%
9	1	0.75	0.5	1.740%	54	0.5	1	1	1.709%	99	0.25	0	0	1.685%
10	1	0.5	0.25	1.740%	55	0.5	0	1	1.709%	100	0.25	0.25	0.25	1.685%
11	1	0	0.5	1.739%	56	0.5	0	0.75	1.708%	101	0.75	1	0.25	1.685%
12	0.75	0.25	0.25	1.736%	57	1	0.25	1	1.708%	102	0.5	0.75	0	1.685%
13	0.75	0.5	0.25	1.736%	58	0.75	0.75	1	1.707%	103	0	0.75	0	1.685%
14	0.25	0.5	0.75	1.736%	59	1	1	1	1.707%	104	0.5	0.5	0.25	1.683%
15	0.75	0.75	0.5	1.734%	60	0	1	0.75	1.707%	105	0.5	0	0.5	1.683%
16	0.5	1	0	1.732%	61	0	0	1	1.706%	106	0.25	0.5	0.5	1.683%
17	0.25	0.75	0.75	1.732%	62	0	0.75	0.5	1.705%	107	0	0.5	0	1.683%
18	0.5	0.75	0.5	1.732%	63	0.75	0.5	0	1.705%	108	0.25	0.75	0	1.682%
19	0.5	0.25	0.75	1.731%	64	1	0.75	0.75	1.704%	109	1	0.75	0	1.679%
20	0.5	1	0.5	1.731%	65	0.25	1	0.75	1.703%	110	0.75	0.5	0.75	1.678%
21	0.75	0	1	1.730%	66	0	0	0.5	1.703%	111	1	1	0.75	1.678%
22	0	0	0.75	1.728%	67	0.5	0.25	1	1.703%	112	0.75	0.5	0.5	1.676%
23	0.25	0	0.75	1.727%	68	1	1	0.25	1.702%	113	0.25	0.25	0.75	1.676%
24	0.75	0	0.75	1.727%	69	0.75	0	0.5	1.702%	114	0.5	0.25	0.5	1.676%
25	0.25	0.5	1	1.726%	70	0	1	0.5	1.702%	115	0	0.75	0.75	1.675%
26	0.75	1	0.5	1.726%	71	1	0.5	0.5	1.701%	116	0	0.25	0	1.674%
27	0	1	0	1.725%	72	0.25	0.25	0	1.701%	117	0.25	0	1	1.673%
28	0	1	1	1.725%	73	0.75	0	0.25	1.701%	118	0.5	0.75	1	1.673%
29	0.75	0.25	0.5	1.724%	74	0	0.25	0.5	1.699%	119	1	0.5	0	1.672%
30	0	0	0.25	1.724%	75	0.25	1	0.5	1.699%	120	0.75	1	0	1.668%
31	0.25	0.5	0.25	1.723%	76	0	0.5	0.75	1.698%	121	0	0.5	0.25	1.668%
32	1	0.25	0.25	1.723%	77	0.5	0	0.25	1.698%	122	0.25	0	0.5	1.661%
33	0.5	0.5	0	1.722%	78	0.75	0.75	0.75	1.698%	123	0.25	0.5	0	1.661%
34	0.75	1	0.75	1.722%	79	0	0.75	1	1.698%	124	0.25	0.25	0.5	1.656%
35	1	1	0	1.721%	80	0.75	0.25	1	1.697%	125	1	0	0.75	1.647%
36	1	0	0.25	1.721%	81	0.75	0.75	0.25	1.697%					
37	0.25	0.75	1	1.720%	82	1	0.5	0.75	1.697%					
38	0.75	0.25	0.75	1.720%	83	0.25	0.75	0.25	1.697%					
39	0	0.75	0.25	1.719%	84	1	1	0.5	1.696%					
40	1	0.75	0.25	1.719%	85	0.25	1	1	1.695%					
41	0.75	0.25	0	1.717%	86	0.5	0.75	0.75	1.692%					
42	1	0	1	1.716%	87	0.5	0.5	0.75	1.692%					
43	0.5	0	0	1.716%	88	1	0.25	0.5	1.691%					
44	0	0.25	0.75	1.715%	89	0.5	1	0.75	1.690%					
45	1	0.25	0	1.714%	90	0	1	0.25	1.690%					

Table D.20: F-measure of the FreqLogit model on a monthly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.25	0.5	1.764%	46	0	0.75	0	1.721%	91	0	1	0	1.700%
2	0.25	0.75	0	1.760%	47	0.5	0	0.5	1.721%	92	1	1	0.25	1.700%
3	0.25	0	0.25	1.754%	48	0.75	0	0.25	1.721%	93	1	1	0.5	1.700%
4	0.5	0.25	0	1.753%	49	0.75	1	0.5	1.720%	94	0.75	1	1	1.700%
5	0.5	0.5	0.25	1.753%	50	0	0	0.25	1.720%	95	0.25	0.25	0.25	1.699%
6	0.75	0.75	0	1.752%	51	0.75	0.25	0.75	1.720%	96	0.5	1	0.75	1.698%
7	1	0.75	0.5	1.746%	52	0	0.5	1	1.720%	97	0.25	0.25	0.75	1.698%
8	0	0	0.75	1.745%	53	0.25	1	0	1.719%	98	1	0.75	0	1.698%
9	0.75	0.5	1	1.744%	54	0.75	0.75	0.5	1.719%	99	1	0.75	0.75	1.697%
10	0.5	0.5	0.75	1.744%	55	1	0.5	0.5	1.718%	100	0.5	0.75	0.25	1.697%
11	1	0	0.5	1.741%	56	0.25	0.25	0.5	1.718%	101	0	0.25	0.25	1.694%
12	0.5	1	1	1.741%	57	0.5	0.25	0.75	1.718%	102	0.75	0.25	0.25	1.694%
13	0	0.75	0.5	1.741%	58	0.5	0	0.75	1.718%	103	0	1	1	1.692%
14	0.75	0	0.5	1.741%	59	0.75	0.25	0	1.717%	104	0	0.5	0.75	1.690%
15	0.25	1	0.25	1.741%	60	0.75	0.25	1	1.717%	105	0.25	0.25	1	1.690%
16	0.75	0.75	0.75	1.741%	61	0.25	0	0.75	1.716%	106	0.75	0.5	0.5	1.689%
17	0.75	0.75	1	1.737%	62	0	0	1	1.715%	107	0.5	0.75	1	1.688%
18	0	0.75	1	1.737%	63	0.5	0	1	1.715%	108	0.25	0	0.5	1.687%
19	1	0.75	0.25	1.736%	64	1	0	0	1.713%	109	0.25	1	0.75	1.686%
20	1	0.25	1	1.735%	65	0	0.25	0	1.713%	110	0.75	1	0	1.686%
21	0.75	0.5	0	1.735%	66	0.25	0.5	0.75	1.713%	111	0.5	0.5	0.5	1.686%
22	1	0.25	0.75	1.735%	67	0	0.75	0.25	1.709%	112	0.5	1	0.5	1.686%
23	0	0.5	0.25	1.734%	68	0.25	0.5	0.25	1.707%	113	1	1	0.75	1.682%
24	0.5	0	0	1.733%	69	0.25	0.75	0.75	1.705%	114	1	1	1	1.682%
25	0	1	0.5	1.732%	70	0	0.5	0	1.704%	115	0.5	1	0	1.681%
26	1	0	1	1.732%	71	0.5	0.75	0.75	1.704%	116	0	1	0.25	1.678%
27	0.25	1	1	1.732%	72	0.25	0.75	0.25	1.704%	117	0.75	0	0	1.677%
28	1	0.25	0.25	1.732%	73	0	0	0.5	1.704%	118	0.5	0.5	1	1.675%
29	0.5	0.5	0	1.729%	74	1	0.5	0.25	1.704%	119	0.75	0.5	0.75	1.674%
30	0	0.75	0.75	1.729%	75	1	1	0	1.703%	120	1	0.5	0.75	1.673%
31	0.75	0.5	0.25	1.728%	76	0.75	1	0.25	1.703%	121	1	0.25	0	1.672%
32	0	0	0	1.728%	77	0.75	1	0.75	1.703%	122	1	0.5	1	1.669%
33	0.25	0.5	0.5	1.727%	78	0.75	0	0.75	1.703%	123	0.25	1	0.5	1.667%
34	1	0.5	0	1.726%	79	0.5	0	0.25	1.703%	124	0.5	1	0.25	1.667%
35	0.75	0.75	0.25	1.726%	80	1	0.75	1	1.702%	125	0.5	0.25	0.25	1.664%
36	0	1	0.75	1.726%	81	0.25	0.5	1	1.702%					
37	0.5	0.75	0	1.726%	82	0.25	0.75	0.5	1.702%					
38	1	0	0.25	1.724%	83	0.5	0.25	0.5	1.701%					
39	1	0.25	0.5	1.724%	84	0	0.25	1	1.701%					
40	0.25	0	0	1.724%	85	0.25	0.5	0	1.701%					
41	0.75	0	1	1.724%	86	1	0	0.75	1.701%					
42	0.5	0.25	1	1.723%	87	0	0.5	0.5	1.701%					
43	0	0.25	0.5	1.723%	88	0	0.25	0.75	1.700%					
44	0.5	0.75	0.5	1.722%	89	0.25	0.75	1	1.700%					
45	0.25	0	1	1.722%	90	0.25	0.25	0	1.700%					

Table D.21: F-measure of the FreqRandom model on a monthly basis, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.5	0	1	7.911%	46	0.5	0.5	0.5	7.816%	91	0	0.25	0.25	7.779%
2	1	0.5	0	7.898%	47	0.25	0.75	0.25	7.816%	92	1	0.25	0.25	7.779%
3	0.75	0	0.75	7.881%	48	0.25	0	0.75	7.815%	93	0.25	0.25	0.25	7.778%
4	0	0.5	0.25	7.879%	49	0.25	0	0	7.815%	94	0.5	0.25	0.75	7.778%
5	0.75	0.25	0.75	7.877%	50	0.25	0	0.25	7.813%	95	1	1	0.25	7.773%
6	0	1	1	7.876%	51	0.25	0.25	1	7.812%	96	0	0.25	0	7.772%
7	1	0	0.75	7.873%	52	0.5	0	0.75	7.810%	97	0.5	0.5	0	7.768%
8	0.75	0	0.25	7.873%	53	0.75	0.75	0.5	7.810%	98	0.5	1	1	7.767%
9	0.75	0.5	0.25	7.872%	54	0.75	0	0.5	7.809%	99	0.25	0.25	0.5	7.767%
10	0.5	0.75	0.5	7.871%	55	0	0.5	1	7.809%	100	0.75	1	0.75	7.766%
11	1	1	1	7.865%	56	0.25	0.75	0.5	7.808%	101	0.5	0	0.5	7.760%
12	0.75	0.5	0.5	7.863%	57	0	1	0	7.807%	102	0.5	0.25	1	7.757%
13	0.5	1	0.5	7.862%	58	1	0.75	0.25	7.807%	103	0	0.75	0	7.753%
14	0	1	0.75	7.860%	59	0	0.25	1	7.806%	104	1	0	0	7.753%
15	0.5	0	0	7.860%	60	1	0.5	0.5	7.805%	105	0	0.25	0.75	7.752%
16	0.75	0.75	0.25	7.859%	61	0	1	0.25	7.805%	106	0	0.5	0.75	7.750%
17	0.5	0.75	0.75	7.856%	62	1	0	1	7.805%	107	0.25	0	0.5	7.747%
18	0.5	1	0	7.855%	63	1	0	0.5	7.804%	108	1	0.5	0.75	7.745%
19	0	0	1	7.854%	64	0.25	0.75	0	7.804%	109	0.75	0.25	1	7.745%
20	0.75	0	0	7.849%	65	1	0.75	0.75	7.804%	110	0.5	0.75	0	7.741%
21	0.25	1	0.75	7.846%	66	0.25	0.5	0.25	7.804%	111	0.75	0	1	7.741%
22	0.75	0.75	0	7.845%	67	0.5	0.5	0.75	7.803%	112	0	0.5	0.5	7.738%
23	0.5	1	0.25	7.843%	68	0.75	0.25	0.5	7.803%	113	1	1	0.5	7.737%
24	0.25	0.75	1	7.842%	69	0.25	0.25	0.75	7.803%	114	0.25	1	0.25	7.736%
25	0.75	0.25	0.25	7.838%	70	0	0.75	0.5	7.802%	115	0	0.75	0.25	7.734%
26	0.5	0.25	0.25	7.837%	71	0.5	0.5	1	7.802%	116	1	1	0.75	7.734%
27	0.5	0.75	0.25	7.834%	72	0.75	0.5	1	7.802%	117	0.75	0.5	0.75	7.734%
28	1	0.75	0.5	7.833%	73	0	0	0.5	7.802%	118	0.75	1	0.25	7.727%
29	0	0.5	0	7.832%	74	0.75	0.25	0	7.800%	119	0.25	0.5	1	7.727%
30	0	0.75	1	7.831%	75	0.5	0.75	1	7.800%	120	1	0.25	0	7.722%
31	0	0	0.25	7.827%	76	1	0	0.25	7.799%	121	0.25	0.75	0.75	7.717%
32	0	0	0.75	7.827%	77	1	0.75	0	7.798%	122	0.5	0.5	0.25	7.709%
33	0.25	0.25	0	7.826%	78	0.75	1	0.5	7.798%	123	1	0.25	0.5	7.704%
34	0	0.75	0.75	7.826%	79	1	0.5	0.25	7.794%	124	0	1	0.5	7.703%
35	0.75	1	1	7.826%	80	0.5	1	0.75	7.794%	125	0.5	0.25	0.5	7.702%
36	0.75	0.75	1	7.825%	81	0.25	1	0.5	7.794%					
37	0.25	0.5	0	7.824%	82	0	0.25	0.5	7.793%					
38	0.25	0.5	0.75	7.824%	83	0.75	0.5	0	7.792%					
39	0	0	0	7.823%	84	1	0.25	0.75	7.792%					
40	1	0.75	1	7.819%	85	0.75	1	0	7.792%					
41	0.25	0	1	7.819%	86	1	0.25	1	7.791%					
42	0.5	0	0.25	7.819%	87	0.25	1	0	7.790%					
43	0.75	0.75	0.75	7.818%	88	1	0.5	1	7.784%					
44	0.25	1	1	7.817%	89	0.5	0.25	0	7.781%					
45	0.25	0.5	0.5	7.817%	90	1	1	0	7.781%					

Table D.22: F-measure of the FreqLCA model over the whole simulation period, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.5	0.25	1	3.624%	46	1	0.25	0.25	3.574%	91	1	1	0.25	3.570%
2	0	0.25	1	3.621%	47	1	0.5	0.25	3.573%	92	0.25	0	0.5	3.569%
3	0	0.25	0.5	3.621%	48	0.5	0.75	0.5	3.572%	93	0	0	0.5	3.569%
4	0.25	0.25	0.75	3.619%	49	0.25	0.5	0.25	3.572%	94	0.5	0	0.5	3.569%
5	0	0.25	0.75	3.619%	50	0.5	0	0.25	3.572%	95	1	1	0	3.569%
6	0.25	0.25	1	3.614%	51	0.5	0.5	0.25	3.572%	96	0.75	0	0	3.569%
7	0.5	0.25	0.5	3.612%	52	0.25	0	0.25	3.572%	97	0.75	0.75	0	3.569%
8	0.5	0.25	0.75	3.605%	53	1	0.5	0	3.572%	98	0	0.75	0.25	3.569%
9	1	0.25	0.75	3.603%	54	1	0.75	0	3.572%	99	1	0	1	3.569%
10	0.75	0.25	0.5	3.600%	55	0.5	1	0	3.572%	100	1	0.75	0.5	3.568%
11	0.75	0.25	0.75	3.600%	56	0.75	0	0.75	3.572%	101	0.75	0	1	3.568%
12	0.5	1	1	3.599%	57	1	0.25	0	3.572%	102	0.75	1	0	3.568%
13	0	0.5	0.75	3.598%	58	0.5	1	0.25	3.571%	103	0.25	0	0	3.568%
14	1	0.25	1	3.598%	59	0.75	1	0.25	3.571%	104	0	0.75	0	3.568%
15	0.25	0.25	0.5	3.596%	60	0	1	0.25	3.571%	105	0.75	0.75	0.5	3.567%
16	0	1	0.75	3.591%	61	0	0	0.25	3.571%	106	1	0.75	0.75	3.567%
17	0.75	0.25	1	3.590%	62	1	0	0.5	3.571%	107	0	0.5	1	3.567%
18	1	0.25	0.5	3.590%	63	0.5	0	1	3.571%	108	0	0.75	0.5	3.567%
19	0	1	1	3.589%	64	0.5	0.25	0	3.571%	109	0.25	0.75	0.25	3.567%
20	0.75	1	1	3.589%	65	0	0.5	0	3.571%	110	0.25	0.75	0.5	3.566%
21	0.5	1	0.75	3.588%	66	0.75	0	0.5	3.571%	111	1	1	1	3.565%
22	0	0.25	0.25	3.588%	67	0.5	0.75	0	3.571%	112	0.25	0.25	0	3.565%
23	0.75	1	0.5	3.587%	68	0	1	0	3.571%	113	0.5	0.5	1	3.564%
24	0.25	1	1	3.586%	69	0.75	0.5	0	3.571%	114	0.75	0.75	0.75	3.564%
25	0.75	1	0.75	3.584%	70	1	0	0.25	3.571%	115	0.75	0.5	1	3.563%
26	1	1	0.75	3.583%	71	1	0	0.75	3.571%	116	0.25	0.5	1	3.563%
27	0.75	0.5	0.5	3.583%	72	0.75	0	0.25	3.571%	117	0	0.75	0.75	3.560%
28	0.25	0.25	0.25	3.581%	73	0.75	0.25	0	3.571%	118	0.5	0.75	0.75	3.560%
29	0	1	0.5	3.581%	74	0	0	1	3.571%	119	0.25	0.75	0.75	3.559%
30	0.25	0.5	0.75	3.581%	75	0.25	0	1	3.571%	120	1	0.5	1	3.554%
31	0.75	0.5	0.75	3.579%	76	0	0.25	0	3.571%	121	0.75	0.75	1	3.549%
32	0.75	0.25	0.25	3.579%	77	0	0	0.75	3.571%	122	0.25	0.75	1	3.546%
33	0.5	1	0.5	3.579%	78	0.25	0	0.75	3.571%	123	1	0.75	1	3.545%
34	0.5	0.5	0.5	3.579%	79	1	0	0	3.571%	124	0.5	0.75	1	3.545%
35	0.5	0.25	0.25	3.578%	80	1	0.75	0.25	3.571%	125	0	0.75	1	3.537%
36	0.25	0.5	0.5	3.578%	81	0	0	0	3.571%					
37	1	0.5	0.5	3.577%	82	0.25	1	0.25	3.570%					
38	1	0.5	0.75	3.577%	83	0.75	0.75	0.25	3.570%					
39	0.25	1	0.75	3.577%	84	0.5	0	0	3.570%					
40	0.5	0.5	0.75	3.576%	85	0.5	0	0.75	3.570%					
41	0.25	1	0.5	3.576%	86	0.5	0.5	0	3.570%					
42	0	0.5	0.5	3.576%	87	0.25	0.75	0	3.570%					
43	0.75	0.5	0.25	3.575%	88	0.25	0.5	0	3.570%					
44	1	1	0.5	3.575%	89	0.25	1	0	3.570%					
45	0	0.5	0.25	3.574%	90	0.5	0.75	0.25	3.570%					

Table D.23: F-measure of the FreqMax model over the whole simulation period, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.25	1	0.25	7.934%	46	0	1	0.75	7.844%	91	0.5	0.5	1	7.805%
2	0	0	0.25	7.907%	47	0.25	0	0.25	7.844%	92	0	0.25	0.5	7.805%
3	0.75	0.25	0.75	7.903%	48	0.25	1	1	7.842%	93	0.25	0	0.5	7.804%
4	0.75	1	0.5	7.900%	49	0	0.5	0	7.841%	94	0.25	0.25	0.25	7.803%
5	0.5	0.5	0.5	7.897%	50	1	1	1	7.841%	95	1	0.5	0.75	7.803%
6	0	0.25	0.75	7.895%	51	1	0.25	0.5	7.840%	96	1	0.25	0.25	7.800%
7	1	0.75	1	7.893%	52	0.25	0.75	0.5	7.839%	97	1	0.75	0	7.800%
8	0	0.25	0.25	7.887%	53	1	1	0.5	7.839%	98	0.75	0.5	0	7.800%
9	0.25	0.5	0.75	7.886%	54	0.25	1	0	7.838%	99	0.25	0	0	7.799%
10	1	0.25	1	7.884%	55	1	1	0	7.837%	100	0	0	0.5	7.797%
11	1	0	0.5	7.884%	56	0.5	0.5	0	7.837%	101	0.75	0.5	0.25	7.796%
12	1	1	0.75	7.882%	57	0.75	0.75	0.5	7.835%	102	0.25	0.5	0	7.794%
13	1	0	0.25	7.882%	58	0.5	1	0.75	7.832%	103	0.25	0.25	0.75	7.794%
14	0.75	0.25	0.25	7.881%	59	0.75	1	0.25	7.831%	104	1	1	0.25	7.792%
15	0.25	0.5	0.25	7.880%	60	0.75	0.5	1	7.830%	105	0	1	0.5	7.792%
16	1	0.5	0.25	7.879%	61	0.25	0.25	0	7.828%	106	0.5	0	1	7.791%
17	0.25	0.75	0	7.877%	62	0.25	1	0.75	7.827%	107	0	0.5	0.25	7.788%
18	0.25	0.75	0.75	7.877%	63	0.75	0.5	0.75	7.825%	108	0.5	0.75	1	7.785%
19	0.25	0	0.75	7.876%	64	0.5	0.5	0.75	7.824%	109	0.75	0.75	0	7.784%
20	0.5	0.25	0.25	7.873%	65	0.5	0.25	0	7.822%	110	0	0.5	0.5	7.783%
21	0	0	1	7.873%	66	0	0.5	0.75	7.822%	111	0.5	0.25	0.75	7.782%
22	0.75	0	0.75	7.872%	67	0	0.75	0.5	7.822%	112	0.5	0	0.75	7.772%
23	1	0.5	1	7.871%	68	0	1	1	7.821%	113	0.75	0	0.5	7.769%
24	1	0.75	0.25	7.869%	69	0.75	0	0	7.820%	114	0.75	0.25	0.5	7.768%
25	1	0	1	7.868%	70	0	0.75	0.25	7.820%	115	1	0.5	0	7.768%
26	0	0.5	1	7.866%	71	0	0	0	7.820%	116	0.5	0	0.5	7.761%
27	0	0.25	1	7.863%	72	0	1	0.25	7.819%	117	0	0.75	0.75	7.760%
28	1	0	0	7.862%	73	0.25	0.75	1	7.819%	118	0.25	0	1	7.758%
29	0.25	0.25	0.5	7.862%	74	0.5	0.5	0.25	7.818%	119	0.75	0.75	0.25	7.755%
30	0.5	0	0	7.859%	75	1	0.75	0.75	7.818%	120	0.5	0.75	0.25	7.755%
31	0.5	0.75	0.5	7.859%	76	0.5	1	0.5	7.817%	121	1	0	0.75	7.752%
32	0.5	1	0	7.857%	77	0.5	0.75	0	7.814%	122	1	0.25	0.75	7.752%
33	0	0.75	0	7.857%	78	0	0.75	1	7.814%	123	0.75	0.5	0.5	7.751%
34	0.75	1	0.75	7.852%	79	1	0.25	0	7.814%	124	0.5	0.25	0.5	7.742%
35	0	1	0	7.851%	80	0.25	1	0.5	7.814%	125	0.75	0.75	0.75	7.708%
36	1	0.5	0.5	7.850%	81	0	0	0.75	7.813%					
37	0.75	0.25	1	7.848%	82	0.75	0.25	0	7.811%					
38	0.5	1	1	7.848%	83	0.5	1	0.25	7.811%					
39	0.75	0	1	7.847%	84	0.5	0.25	1	7.811%					
40	0.25	0.25	1	7.846%	85	0	0.25	0	7.810%					
41	0.25	0.5	0.5	7.845%	86	0.75	0	0.25	7.810%					
42	0.25	0.5	1	7.844%	87	0.5	0	0.25	7.810%					
43	0.75	1	0	7.844%	88	0.75	0.75	1	7.809%					
44	0.25	0.75	0.25	7.844%	89	0.5	0.75	0.75	7.808%					
45	1	0.75	0.5	7.844%	90	0.75	1	1	7.805%					

Table D.24: F-measure of the FreqLogit model over the whole simulation period, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.75	0.75	7.956%	46	0.5	0.25	0.75	7.854%	91	1	0.5	0	7.819%
2	0.75	0.75	0.75	7.943%	47	0	0.25	0.75	7.852%	92	0	0.25	0	7.819%
3	0.5	0.75	0.75	7.930%	48	0.5	1	0	7.852%	93	0	0.25	0.5	7.818%
4	0.75	0.75	0	7.925%	49	0.25	1	0.25	7.852%	94	0.5	0.5	1	7.818%
5	0.75	1	1	7.924%	50	0.25	0	1	7.850%	95	1	0.75	0	7.817%
6	0.75	0.75	1	7.918%	51	0.75	0.5	0.5	7.849%	96	0.25	0.25	0.25	7.816%
7	1	0.25	1	7.913%	52	0.5	0.75	0	7.849%	97	0.5	0.5	0.5	7.815%
8	1	0	0.5	7.903%	53	0.5	0.5	0.25	7.848%	98	1	1	0.5	7.814%
9	0.5	1	1	7.902%	54	1	0.75	0.75	7.847%	99	1	0	1	7.810%
10	0.75	0.5	0.25	7.901%	55	0.75	0.25	0.75	7.846%	100	0	0.5	0.25	7.809%
11	1	0	0	7.901%	56	0.5	0.5	0.75	7.845%	101	0	0	0.25	7.808%
12	0.25	1	1	7.900%	57	0.25	1	0	7.845%	102	0.75	0	0.25	7.808%
13	1	0.5	1	7.898%	58	0	1	0	7.843%	103	1	0.75	0.5	7.807%
14	0	0	0.75	7.895%	59	0	0.5	1	7.843%	104	0	0.25	0.25	7.807%
15	1	1	1	7.894%	60	1	0.5	0.5	7.843%	105	1	0.5	0.75	7.807%
16	0	0	0	7.891%	61	1	0.25	0.5	7.843%	106	0.5	0.5	0	7.806%
17	1	1	0.75	7.890%	62	0.5	0.75	0.5	7.842%	107	0.25	0.25	0	7.802%
18	0.75	0.75	0.5	7.888%	63	0	0.75	0	7.841%	108	0.75	0.5	0	7.802%
19	0.25	0.5	0.25	7.886%	64	0	0	0.5	7.841%	109	0.5	0	0.5	7.799%
20	0.75	0	0.75	7.882%	65	1	1	0.25	7.840%	110	0	0.75	0.25	7.797%
21	1	0.75	0.25	7.882%	66	0.25	0.5	0.5	7.839%	111	0.5	0	0.75	7.796%
22	0.25	0.75	0.25	7.879%	67	0.75	1	0.5	7.838%	112	0.25	0.5	0	7.795%
23	1	0.25	0.75	7.876%	68	0	1	0.5	7.838%	113	0.5	0	0.25	7.794%
24	1	0	0.25	7.876%	69	1	0.25	0	7.837%	114	0.75	0.25	0	7.790%
25	0.25	0.75	0	7.874%	70	0.5	0.25	0	7.835%	115	0.75	0.25	0.25	7.789%
26	0.75	0.5	1	7.872%	71	0	0.25	1	7.835%	116	0	0	1	7.789%
27	1	1	0	7.869%	72	0.25	1	0.75	7.835%	117	0	1	0.25	7.789%
28	0.75	1	0.25	7.868%	73	0	0.5	0.75	7.834%	118	0.75	0	1	7.787%
29	0.75	0.25	0.5	7.868%	74	1	0.25	0.25	7.834%	119	0.5	0.25	0.5	7.782%
30	0.5	0.75	1	7.867%	75	0	0.75	0.5	7.834%	120	0.5	0.75	0.25	7.779%
31	0.25	0.75	0.75	7.865%	76	0.25	0.75	0.5	7.833%	121	0	1	1	7.778%
32	0.75	0	0.5	7.865%	77	0.75	0.75	0.25	7.832%	122	0.25	0.5	0.75	7.776%
33	0	1	0.75	7.864%	78	0	0.5	0.5	7.830%	123	0.5	1	0.5	7.774%
34	0.25	0	0.25	7.863%	79	1	0	0.75	7.830%	124	0.5	0.25	0.25	7.761%
35	0.75	1	0	7.863%	80	0.25	0.25	1	7.830%	125	0.75	0	0	7.761%
36	0.25	0.25	0.75	7.862%	81	1	0.5	0.25	7.828%					
37	0.75	0.5	0.75	7.861%	82	0.5	1	0.25	7.828%					
38	0.25	0.5	1	7.860%	83	0.25	0	0.75	7.826%					
39	0.75	1	0.75	7.858%	84	0.25	0.25	0.5	7.825%					
40	0.5	0.25	1	7.857%	85	0.25	0	0.5	7.825%					
41	0	0.5	0	7.857%	86	0	0.75	1	7.824%					
42	0.5	1	0.75	7.856%	87	0.75	0.25	1	7.824%					
43	1	0.75	1	7.856%	88	0.25	0	0	7.823%					
44	0.25	0.75	1	7.856%	89	0.25	1	0.5	7.821%					
45	0.5	0	0	7.854%	90	0.5	0	1	7.819%					

Table D.25: F-measure of the FreqRandom model over the whole simulation period, for the creator topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	1	0.25	0.519%	46	0.25	0.25	0.5	0.493%	91	1	0.25	1	0.483%
2	0	0.25	0.25	0.513%	47	0.25	0.5	0.25	0.492%	92	0.75	1	0	0.483%
3	1	1	1	0.509%	48	0	0.5	0.75	0.492%	93	1	0.25	0.75	0.482%
4	0.5	1	1	0.508%	49	0	0.75	0.5	0.492%	94	0.5	0	0.5	0.482%
5	0.75	0.5	0	0.508%	50	0.75	0.25	1	0.492%	95	0	0	0	0.482%
6	0.75	0.75	0.75	0.507%	51	0	0.75	0.75	0.492%	96	0	0.25	0.5	0.482%
7	0	0.25	0	0.506%	52	0	1	0.75	0.492%	97	0.25	0	0.5	0.482%
8	1	0	0.5	0.506%	53	0	0	0.25	0.491%	98	0	0.5	0.5	0.481%
9	0.75	0.5	0.25	0.506%	54	1	0.25	0.5	0.491%	99	0.25	0.5	0.5	0.481%
10	0.5	1	0	0.505%	55	0.5	1	0.75	0.491%	100	0.75	1	0.25	0.481%
11	0.5	0.75	1	0.504%	56	1	1	0.25	0.491%	101	1	0.5	0.25	0.481%
12	1	0.5	0.5	0.504%	57	0.75	0.75	0	0.490%	102	0.25	0.25	0.25	0.479%
13	1	1	0	0.504%	58	0.75	0	1	0.490%	103	0	0.25	0.75	0.479%
14	0.25	0.75	1	0.503%	59	0.5	0	0.25	0.490%	104	0.5	0	0.75	0.479%
15	0	0.75	1	0.503%	60	0.75	0	0	0.490%	105	0.75	1	0.5	0.479%
16	1	0.75	0	0.502%	61	0.75	0.75	0.5	0.490%	106	0.75	1	1	0.478%
17	0.25	1	0.5	0.502%	62	0.75	0.25	0.5	0.489%	107	0.5	0.5	0.5	0.478%
18	0.5	0.25	0.75	0.501%	63	0.75	0.75	1	0.489%	108	0.5	0.25	0	0.477%
19	0.75	0.25	0.75	0.501%	64	0.25	1	0	0.489%	109	0	0	0.5	0.477%
20	0	1	0.5	0.501%	65	0.75	0	0.75	0.488%	110	0.25	0.75	0.25	0.477%
21	0.5	0.5	0.75	0.501%	66	0	0.5	0	0.488%	111	0.5	0.25	0.25	0.477%
22	0.5	0.75	0.25	0.500%	67	1	0.5	1	0.488%	112	0	0.5	0.25	0.476%
23	0.5	0.75	0.75	0.500%	68	0.25	0.5	0.75	0.488%	113	1	1	0.5	0.476%
24	0	1	0	0.500%	69	0.25	1	0.75	0.488%	114	0.25	0.25	1	0.476%
25	1	1	0.75	0.500%	70	1	0.5	0	0.488%	115	0.5	0	0	0.476%
26	0.25	1	1	0.499%	71	0.25	0.25	0.75	0.488%	116	1	0	1	0.476%
27	0	0.75	0	0.499%	72	0	0.25	1	0.488%	117	0.75	0	0.25	0.476%
28	0.75	0	0.5	0.498%	73	0.5	0.75	0	0.487%	118	1	0	0.25	0.476%
29	0.25	0	0.25	0.498%	74	0.25	0.25	0	0.487%	119	0.25	0.5	1	0.476%
30	0.5	0	1	0.498%	75	1	0.75	0.25	0.487%	120	0.25	0.75	0.75	0.475%
31	1	0.75	0.75	0.498%	76	0.75	0.25	0	0.487%	121	0.75	0.75	0.25	0.474%
32	0.25	1	0.25	0.498%	77	0.75	0.25	0.25	0.487%	122	1	0.75	1	0.473%
33	0.75	0.5	0.75	0.497%	78	1	0.25	0.25	0.486%	123	0.5	1	0.25	0.471%
34	0.5	0.25	1	0.497%	79	0.75	1	0.75	0.485%	124	0.5	0.5	0.25	0.465%
35	0.5	0.25	0.5	0.497%	80	0.5	0.75	0.5	0.485%	125	1	0	0	NaN
36	0.75	0.5	0.5	0.497%	81	0	0	0.75	0.485%					
37	0.25	0.75	0.5	0.496%	82	0.75	0.5	1	0.485%					
38	0.25	0.5	0	0.496%	83	0	0	1	0.485%					
39	0.5	0.5	0	0.496%	84	0	0.5	1	0.484%					
40	0.5	1	0.5	0.495%	85	1	0.5	0.75	0.484%					
41	0	1	1	0.495%	86	0.5	0.5	1	0.484%					
42	1	0.25	0	0.494%	87	1	0	0.75	0.484%					
43	0.25	0	0	0.494%	88	0	0.75	0.25	0.484%					
44	1	0.75	0.5	0.493%	89	0.25	0	0.75	0.483%					
45	0.25	0.75	0	0.493%	90	0.25	0	1	0.483%					

Table D.26: F-measure of the FreqLCA model on a weekly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.75	0	0.669%	46	0.5	1	0	0.639%	91	0.5	0.25	1	0.625%
2	0	1	0	0.666%	47	0	0.25	1	0.639%	92	0.5	0.75	0.25	0.624%
3	0.25	0	0.5	0.659%	48	0.5	1	0.5	0.638%	93	0	0.25	0	0.624%
4	1	0	0.5	0.658%	49	0.25	0.75	1	0.636%	94	0.5	1	0.25	0.624%
5	0.5	0	0	0.656%	50	0.25	0	0.25	0.636%	95	1	0.25	1	0.623%
6	0.75	1	1	0.655%	51	0.75	0.5	0.5	0.636%	96	0.75	0.5	1	0.623%
7	0.5	0.5	0	0.654%	52	1	0.25	0.25	0.636%	97	0.75	0.75	0.5	0.623%
8	1	0.25	0	0.654%	53	0.5	0.25	0.25	0.635%	98	1	0.5	1	0.622%
9	1	0	0.25	0.653%	54	0.25	1	0.75	0.635%	99	0.75	1	0	0.622%
10	0.25	0.5	0.25	0.651%	55	0.75	0	0.5	0.635%	100	0	0.25	0.75	0.622%
11	1	0.75	0	0.650%	56	0.25	0.25	0.75	0.635%	101	0.25	0.25	0.5	0.622%
12	0.5	0.75	0.75	0.650%	57	0.5	0.25	0.5	0.635%	102	0.75	0.25	0	0.622%
13	0.25	0.25	0.25	0.649%	58	0.25	0	1	0.635%	103	0.75	1	0.5	0.621%
14	0.25	0.75	0	0.648%	59	0	0.75	1	0.634%	104	0.5	0.25	0	0.620%
15	0.25	0.75	0.25	0.648%	60	0.5	1	1	0.634%	105	0.25	0	0	0.620%
16	0.5	0.5	0.25	0.648%	61	0	0	0.5	0.634%	106	1	0.75	0.25	0.619%
17	0.25	0.5	0.75	0.647%	62	0.25	1	0.25	0.633%	107	0.75	1	0.25	0.619%
18	0.75	0.25	0.25	0.647%	63	1	0.5	0.5	0.633%	108	1	1	1	0.619%
19	0	0.25	0.5	0.646%	64	1	0.75	1	0.633%	109	0.5	1	0.75	0.618%
20	1	1	0.75	0.646%	65	1	0	0	0.633%	110	0	0	0.25	0.618%
21	0.75	0.5	0.25	0.646%	66	1	0.25	0.5	0.633%	111	0	0	0.75	0.618%
22	0.75	0.75	0.75	0.645%	67	0.5	0	0.25	0.633%	112	0	0.5	0.5	0.615%
23	1	0	1	0.645%	68	0.25	0.5	0.5	0.633%	113	0.25	0.25	1	0.615%
24	1	1	0.25	0.645%	69	0.25	0.25	0	0.632%	114	0.25	1	0.5	0.615%
25	0	0.75	0.5	0.644%	70	1	0.25	0.75	0.632%	115	1	0.5	0.75	0.614%
26	0	0.5	1	0.644%	71	0.75	0.25	1	0.631%	116	0.5	0.75	0.5	0.614%
27	0.25	0.5	0	0.643%	72	0.5	0.5	0.5	0.630%	117	0.5	0.5	0.75	0.614%
28	0.75	0.25	0.75	0.642%	73	0	1	0.75	0.630%	118	0.5	0.75	1	0.613%
29	0	0.5	0	0.642%	74	0.5	0	0.5	0.630%	119	0.75	0.25	0.5	0.612%
30	0.25	1	1	0.642%	75	1	1	0.5	0.629%	120	1	0.5	0.25	0.605%
31	0	0.5	0.25	0.642%	76	0.75	0.5	0	0.629%	121	0	0.25	0.25	0.604%
32	1	0	0.75	0.641%	77	0.75	0.75	0	0.629%	122	0	1	1	0.602%
33	0	0.75	0.25	0.641%	78	0.25	0	0.75	0.629%	123	0.25	0.5	1	0.597%
34	1	1	0	0.640%	79	1	0.5	0	0.629%	124	0	1	0.25	NaN
35	0.5	0.75	0	0.640%	80	0	0.75	0.75	0.629%	125	0.25	1	0	NaN
36	0.5	0	0.75	0.640%	81	0.25	0.75	0.5	0.628%					
37	0.75	0.75	1	0.640%	82	0.75	1	0.75	0.627%					
38	0.75	0	0	0.640%	83	1	0.75	0.75	0.627%					
39	0	0	1	0.640%	84	0.75	0.75	0.25	0.626%					
40	0.75	0	0.25	0.640%	85	0.25	0.75	0.75	0.626%					
41	0.75	0	0.75	0.639%	86	0.5	0	1	0.626%					
42	0.75	0.5	0.75	0.639%	87	0	1	0.5	0.626%					
43	0.5	0.5	1	0.639%	88	0.75	0	1	0.626%					
44	0	0	0	0.639%	89	1	0.75	0.5	0.625%					
45	0	0.5	0.75	0.639%	90	0.5	0.25	0.75	0.625%					

Table D.27: F-measure of the FreqMax model on a weekly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0	0.25	0.530%	46	0.25	1	0	0.502%	91	0.25	0.75	0.25	0.492%
2	0	0.5	0	0.529%	47	0	0.75	0.75	0.502%	92	0	0.75	0.5	0.492%
3	0.75	1	0.25	0.528%	48	1	0.75	0.25	0.502%	93	0	1	0.75	0.492%
4	0	1	1	0.524%	49	0.25	1	0.75	0.502%	94	0.5	0.75	0.5	0.491%
5	0.75	0	0.5	0.522%	50	0	0.5	0.25	0.501%	95	0.5	0	0	0.491%
6	1	0	1	0.520%	51	0.25	0	0.25	0.501%	96	0.75	0.75	0	0.491%
7	0.75	0.5	0.75	0.518%	52	0.5	1	0.25	0.501%	97	0.5	0.75	1	0.491%
8	0.75	0.5	1	0.517%	53	0.25	0.25	0	0.501%	98	0.75	0.75	0.5	0.491%
9	0.75	0.75	0.75	0.517%	54	0.25	0	0	0.501%	99	1	0.75	1	0.490%
10	0.25	0.5	0.25	0.515%	55	0	0.25	0.75	0.501%	100	1	0.25	0.5	0.490%
11	0.25	0.5	0.5	0.515%	56	0	0.25	0.25	0.501%	101	1	0.5	0.5	0.489%
12	0	0.5	0.5	0.515%	57	0	0	0.5	0.500%	102	0.25	0.75	0.75	0.489%
13	1	0	0.75	0.515%	58	0	0.25	0	0.500%	103	1	0	0.5	0.489%
14	1	0.75	0.75	0.514%	59	0.75	0.5	0	0.500%	104	0.5	0.5	1	0.488%
15	0.25	0.75	0	0.514%	60	0.25	0	0.75	0.500%	105	0.5	0.25	0.5	0.488%
16	0.75	1	0.5	0.513%	61	0	0	0	0.500%	106	0.75	0	0	0.488%
17	0.25	0.25	1	0.513%	62	0.25	0.25	0.5	0.499%	107	0.5	0	0.25	0.487%
18	0.25	0.75	0.5	0.512%	63	0.75	1	0.75	0.499%	108	0.5	0	0.75	0.487%
19	1	0.75	0	0.512%	64	0.75	0.25	0.75	0.499%	109	0.75	0.25	0.5	0.487%
20	0.25	1	0.25	0.511%	65	0.25	0	0.5	0.499%	110	0.5	0.5	0	0.486%
21	0.75	0.25	0	0.511%	66	0.75	0	1	0.498%	111	1	0.5	0	0.485%
22	0.25	0.75	1	0.509%	67	0.75	0.75	1	0.498%	112	0.5	1	1	0.485%
23	1	1	0	0.508%	68	0.5	0.25	1	0.498%	113	0	0.25	0.5	0.485%
24	0.75	0	0.75	0.508%	69	1	1	0.75	0.497%	114	0.5	0	0.5	0.485%
25	0.75	0.25	1	0.507%	70	0	1	0	0.497%	115	0.75	0.5	0.5	0.484%
26	0	0	0.25	0.507%	71	0.75	1	1	0.497%	116	0	0	0.75	0.484%
27	0.5	0.25	0.75	0.507%	72	0.5	0.25	0.25	0.497%	117	0.5	0.5	0.75	0.483%
28	0.25	1	1	0.506%	73	0.25	0	1	0.497%	118	0.75	0.75	0.25	0.481%
29	1	0.25	1	0.506%	74	0.75	0.25	0.25	0.497%	119	0.5	0.25	0	0.481%
30	1	0	0.25	0.506%	75	1	0.75	0.5	0.496%	120	1	0.5	0.25	0.480%
31	0.75	1	0	0.506%	76	0.25	0.25	0.25	0.496%	121	0.75	0.5	0.25	0.480%
32	1	0	0	0.505%	77	1	1	0.25	0.496%	122	1	0.5	1	0.466%
33	0.25	0.5	0.75	0.505%	78	0.5	1	0.5	0.496%	123	0	1	0.25	NaN
34	0.5	1	0.75	0.505%	79	0.5	0.5	0.5	0.496%	124	0.5	0.5	0.25	NaN
35	0	0.75	0.25	0.505%	80	1	0.25	0	0.495%	125	1	1	1	NaN
36	0.25	0.5	1	0.505%	81	1	0.25	0.25	0.495%					
37	0	0.75	1	0.505%	82	1	0.5	0.75	0.495%					
38	0	0.75	0	0.504%	83	1	0.25	0.75	0.495%					
39	0.25	0.25	0.75	0.504%	84	0	1	0.5	0.494%					
40	0.25	1	0.5	0.504%	85	0.5	0.75	0.75	0.494%					
41	0.5	0	1	0.504%	86	1	1	0.5	0.494%					
42	0	0.5	0.75	0.504%	87	0	0.5	1	0.494%					
43	0.5	0.75	0	0.503%	88	0	0	1	0.494%					
44	0.25	0.5	0	0.503%	89	0	0.25	1	0.493%					
45	0.5	0.75	0.25	0.503%	90	0.5	1	0	0.493%					

Table D.28: F-measure of the FreqLogit model on a weekly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.25	1	0.5	0.525%	46	0.75	0	0.5	0.503%	91	0.5	1	0	0.493%
2	0.75	0	0	0.525%	47	0.25	0	0.75	0.502%	92	0.75	0.5	1	0.492%
3	1	0.75	0.25	0.524%	48	1	0.25	0	0.502%	93	0.5	0.25	0.75	0.492%
4	0.5	0	1	0.522%	49	0.5	0.25	1	0.502%	94	0.5	0.75	1	0.492%
5	1	0.5	0.5	0.520%	50	0.75	0.75	1	0.502%	95	0	0.75	0	0.491%
6	0.75	0.25	1	0.519%	51	1	1	1	0.501%	96	1	0	1	0.491%
7	0.25	0.5	0.5	0.518%	52	1	0.75	0.75	0.501%	97	0	0.75	0.25	0.491%
8	0.75	0.75	0.75	0.517%	53	0.75	0.5	0	0.501%	98	0.5	1	0.25	0.491%
9	0.75	1	0.25	0.516%	54	1	0.5	0.25	0.501%	99	0.75	0	0.25	0.491%
10	1	0	0	0.515%	55	0.5	0.5	1	0.501%	100	0.25	1	1	0.490%
11	0.75	1	1	0.515%	56	0	0	0.5	0.500%	101	0.5	0.25	0.25	0.490%
12	0.75	0.75	0.5	0.515%	57	0	0.25	0	0.500%	102	0.5	1	1	0.490%
13	0.25	0.25	1	0.514%	58	0.25	0	0	0.500%	103	0.75	0.25	0.25	0.489%
14	0	1	0	0.510%	59	0.5	0.5	0.5	0.499%	104	0.5	0	0.25	0.488%
15	0.25	0	0.5	0.510%	60	0	1	0.25	0.499%	105	0.5	0.5	0.75	0.488%
16	0.25	0.5	1	0.510%	61	0	0	0	0.499%	106	0	0	0.25	0.488%
17	1	0	0.5	0.509%	62	1	0.25	0.25	0.499%	107	0.75	0	0.75	0.488%
18	0.75	0.25	0.5	0.509%	63	1	1	0	0.498%	108	0.25	0.25	0.25	0.487%
19	0.75	0.75	0.25	0.509%	64	0.75	1	0.5	0.498%	109	1	0.25	0.75	0.487%
20	1	0.25	1	0.509%	65	0.25	0.5	0.75	0.498%	110	0.75	0.75	0	0.486%
21	0	0.5	0	0.509%	66	0.75	0.25	0.75	0.498%	111	0	0.5	0.75	0.486%
22	0.75	0.5	0.75	0.509%	67	0	0.75	0.5	0.498%	112	0.5	0.25	0	0.486%
23	0	0.25	1	0.509%	68	0.25	0.75	1	0.498%	113	1	0.5	0.75	0.486%
24	0.25	0.75	0	0.508%	69	1	0.75	0	0.498%	114	0.25	0.25	0.5	0.485%
25	0.25	0.25	0.75	0.508%	70	0	0.75	1	0.497%	115	0.25	1	0.75	0.485%
26	0.25	0.5	0	0.508%	71	0	0.5	0.25	0.497%	116	0	1	1	0.484%
27	0.5	0.25	0.5	0.507%	72	1	0.5	0	0.497%	117	0.25	0	0.25	0.482%
28	0.5	0.75	0	0.507%	73	0	0.25	0.5	0.497%	118	0.5	1	0.5	0.481%
29	1	0.5	1	0.507%	74	0.5	1	0.75	0.497%	119	0	0.5	0.5	0.480%
30	0	0.5	1	0.507%	75	0.5	0	0.5	0.497%	120	0.25	0.75	0.75	0.477%
31	0	1	0.75	0.506%	76	1	0	0.75	0.497%	121	0.5	0.75	0.75	0.476%
32	0.75	0.25	0	0.506%	77	1	0.75	1	0.496%	122	0.25	0.75	0.25	0.472%
33	0.5	0.5	0	0.505%	78	0	0.25	0.25	0.496%	123	1	1	0.75	0.471%
34	1	0	0.25	0.505%	79	0	0.75	0.75	0.495%	124	0	0.25	0.75	NaN
35	0.75	0.5	0.25	0.505%	80	0.25	1	0.25	0.495%	125	0.75	1	0	NaN
36	0.5	0.75	0.25	0.505%	81	0.25	0.25	0	0.495%					
37	0	0	1	0.505%	82	0	0	0.75	0.495%					
38	0.5	0.75	0.5	0.504%	83	0.75	1	0.75	0.495%					
39	0.25	0.75	0.5	0.504%	84	1	0.25	0.5	0.495%					
40	0.5	0.5	0.25	0.504%	85	1	1	0.5	0.494%					
41	1	1	0.25	0.504%	86	0	1	0.5	0.494%					
42	0.25	0.5	0.25	0.504%	87	0.75	0	1	0.494%					
43	0.75	0.5	0.5	0.503%	88	0.25	1	0	0.493%					
44	0.25	0	1	0.503%	89	0.5	0	0	0.493%					
45	0.5	0	0.75	0.503%	90	1	0.75	0.5	0.493%					

Table D.29: F-measure of the FreqRandom model on a weekly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	0.5	0.5	1.487%	46	0.25	0.25	0.5	1.436%	91	0.75	1	0.5	1.421%
2	0.5	1	0.5	1.473%	47	0.5	1	0.75	1.435%	92	0	0	0.75	1.420%
3	1	1	1	1.471%	48	0	0.5	1	1.435%	93	0	0.75	0.75	1.420%
4	0	1	1	1.469%	49	0	1	0	1.435%	94	0.75	0	1	1.418%
5	0.75	0.5	0.5	1.469%	50	0.25	1	0.5	1.435%	95	0.25	0.75	0.75	1.418%
6	0	1	0.25	1.468%	51	1	0.25	1	1.435%	96	1	0	0.75	1.417%
7	0.5	0	0.25	1.465%	52	0.75	0.75	1	1.434%	97	1	0.25	0.25	1.416%
8	1	0.25	0	1.459%	53	1	0.5	0	1.434%	98	0.5	0.25	0.25	1.416%
9	0.5	1	0	1.458%	54	0.75	0.25	0.75	1.434%	99	0.75	1	0.75	1.416%
10	0.5	0.5	0.75	1.458%	55	0.5	0.25	0.75	1.433%	100	0	0.25	1	1.415%
11	0.75	0.75	0.75	1.456%	56	0.25	0	0.75	1.432%	101	0	0.75	0.25	1.415%
12	0.25	1	0	1.453%	57	0	0	0.5	1.432%	102	0.25	0	0	1.415%
13	0.75	0	0	1.453%	58	0	0.75	1	1.432%	103	1	0	0.25	1.414%
14	0	0.25	0	1.451%	59	0	0.25	0.25	1.432%	104	0.75	1	0.25	1.414%
15	0.75	0	0.5	1.451%	60	0.5	0.5	0	1.432%	105	1	1	0.25	1.413%
16	0.75	0.25	0.25	1.450%	61	0.25	0.25	0.25	1.432%	106	0.75	0.75	0.25	1.410%
17	0	1	0.75	1.450%	62	0.75	1	1	1.432%	107	0.25	0	1	1.410%
18	0.25	1	1	1.450%	63	0.75	0.5	0.75	1.431%	108	0.75	0.75	0	1.410%
19	0.75	0.5	0.25	1.449%	64	1	0	0.5	1.431%	109	0.25	0.5	1	1.409%
20	1	0.5	0.25	1.449%	65	0	0.75	0.5	1.430%	110	1	0.75	1	1.408%
21	1	0	0	1.449%	66	1	0.25	0.75	1.430%	111	1	0.75	0	1.407%
22	0.25	0.5	0.25	1.449%	67	0.5	0.25	1	1.429%	112	0.75	0.25	0	1.407%
23	0.25	0.75	0	1.448%	68	0.25	0.75	0.25	1.429%	113	1	0.75	0.25	1.407%
24	0.5	0.75	0.25	1.448%	69	0.75	0.5	1	1.428%	114	0.25	0	0.5	1.405%
25	0.5	0	0	1.447%	70	0.25	0.75	0.5	1.428%	115	0.5	0.5	0.25	1.405%
26	0	0.5	0.75	1.445%	71	1	1	0.75	1.428%	116	0.25	0.5	0.75	1.402%
27	0.25	0	0.25	1.444%	72	1	1	0	1.428%	117	0	0.75	0	1.401%
28	0.5	0	0.5	1.443%	73	0.5	0.75	0.5	1.428%	118	0.75	0	0.25	1.400%
29	0.75	0.5	0	1.443%	74	0	0.25	0.5	1.427%	119	0.25	0.25	0.75	1.400%
30	0.25	0.5	0.5	1.443%	75	0.5	1	1	1.427%	120	0.5	0.25	0	1.400%
31	0	0	0.25	1.442%	76	0.25	0.25	0	1.427%	121	1	0	1	1.400%
32	0	0.5	0	1.442%	77	0.75	0.75	0.5	1.427%	122	0	0	0	1.396%
33	0.5	0.75	1	1.442%	78	0	0.5	0.5	1.427%	123	0.5	0.25	0.5	1.395%
34	0.5	0.5	0.5	1.441%	79	0.25	1	0.25	1.426%	124	1	1	0.5	1.392%
35	0	0.25	0.75	1.441%	80	0.5	0.75	0.75	1.425%	125	0.5	0	0.75	1.390%
36	0.25	0.25	1	1.441%	81	0	0.5	0.25	1.424%					
37	0.75	0	0.75	1.440%	82	0.5	0.75	0	1.424%					
38	0.75	1	0	1.439%	83	0.25	0.75	1	1.423%					
39	0.75	0.25	0.5	1.439%	84	0.5	0.5	1	1.423%					
40	1	0.25	0.5	1.439%	85	0	0	1	1.423%					
41	0	1	0.5	1.438%	86	1	0.5	0.75	1.422%					
42	1	0.75	0.75	1.438%	87	0.25	0.5	0	1.422%					
43	1	0.5	1	1.437%	88	0.75	0.25	1	1.421%					
44	0.5	0	1	1.437%	89	0.5	1	0.25	1.421%					
45	1	0.75	0.5	1.436%	90	0.25	1	0.75	1.421%					

Table D.30: F-measure of the FreqLCA model on a monthly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.5	0.5	0.25	2.065%	46	0.75	0	0.25	1.998%	91	0.75	0	0.5	1.973%
2	0.25	0.75	0.5	2.048%	47	0.75	1	1	1.998%	92	1	0	0.5	1.971%
3	1	0.75	0.5	2.048%	48	0.25	0.5	0.75	1.998%	93	1	0	0.25	1.970%
4	0.5	0.25	0.25	2.046%	49	0.5	1	0.5	1.997%	94	0	1	1	1.969%
5	0	1	0.75	2.045%	50	0.25	1	0.25	1.997%	95	0	0.75	1	1.969%
6	0.5	0	0.25	2.042%	51	0.5	0	0.75	1.996%	96	0.75	1	0.75	1.969%
7	0.25	0.25	0	2.036%	52	1	0.75	0	1.995%	97	0	0.25	1	1.968%
8	0.25	0.25	0.25	2.036%	53	0	1	0.25	1.995%	98	0.5	1	0.75	1.966%
9	0	0.75	0.5	2.035%	54	0.25	0	0.25	1.995%	99	0.75	1	0.5	1.966%
10	0	0.75	0.75	2.031%	55	0.5	0.75	0.75	1.994%	100	0.5	0.75	0.5	1.966%
11	0	0	0.5	2.031%	56	0.75	0.5	0.5	1.993%	101	1	0.5	0.5	1.965%
12	1	0	0.75	2.030%	57	1	1	0	1.993%	102	0	0.5	0.5	1.964%
13	0.5	1	0	2.027%	58	1	1	0.25	1.992%	103	0.75	0.75	0	1.962%
14	0.5	0.75	1	2.026%	59	0.75	0.25	1	1.991%	104	0.75	0.75	0.5	1.960%
15	0	1	0	2.024%	60	0.25	1	1	1.991%	105	0.25	0.25	1	1.958%
16	0.5	0.25	0.5	2.023%	61	0.75	1	0.25	1.990%	106	0.25	0.25	0.75	1.955%
17	0.25	1	0.75	2.022%	62	0.75	0.5	0.25	1.989%	107	0.25	1	0	1.955%
18	1	0	0	2.020%	63	0.75	0.25	0.5	1.989%	108	0.5	1	0.25	1.954%
19	0.5	0	1	2.019%	64	1	0.25	0	1.988%	109	0.75	0.5	0.75	1.951%
20	0.75	0.75	0.25	2.019%	65	0.5	0.75	0.25	1.987%	110	0.5	1	1	1.951%
21	1	0	1	2.019%	66	0	1	0.5	1.987%	111	0.5	0.5	0	1.950%
22	1	0.25	0.25	2.019%	67	1	0.75	1	1.986%	112	0	0.25	0.75	1.949%
23	0	0.25	0	2.018%	68	0	0	0	1.985%	113	0.5	0.25	0.75	1.946%
24	0.75	0.5	0	2.017%	69	0	0.5	0.75	1.985%	114	1	0.5	1	1.946%
25	0.75	0	0.75	2.016%	70	0	0	1	1.984%	115	1	1	0.5	1.941%
26	0	0.75	0	2.016%	71	0.25	0.75	0.75	1.984%	116	0.25	1	0.5	1.941%
27	0.25	0.75	0.25	2.015%	72	1	0.5	0	1.984%	117	0.5	0.5	0.75	1.937%
28	0.5	0.5	0.5	2.014%	73	0.25	0	0.75	1.982%	118	0.5	0.25	0	1.936%
29	0.25	0	0.5	2.014%	74	0.25	0.5	0	1.982%	119	1	0.25	0.5	1.932%
30	0.75	0.25	0.25	2.014%	75	0	0	0.75	1.980%	120	0	0.25	0.5	1.932%
31	0.75	0	0	2.013%	76	0	0.25	0.25	1.979%	121	0.75	0.5	1	1.931%
32	0	0.5	0	2.012%	77	0.25	0.25	0.5	1.979%	122	1	1	1	1.931%
33	0.5	0	0.5	2.012%	78	0.75	0.75	1	1.977%	123	1	0.5	0.75	1.926%
34	0.75	1	0	2.011%	79	0.75	0.25	0	1.977%	124	0.5	0.25	1	1.916%
35	1	1	0.75	2.010%	80	0.25	0.75	1	1.977%	125	0.25	0.5	1	1.909%
36	0.25	0.5	0.5	2.009%	81	1	0.75	0.75	1.977%					
37	0	0.5	0.25	2.007%	82	1	0.5	0.25	1.977%					
38	0	0	0.25	2.006%	83	1	0.25	0.75	1.977%					
39	0.5	0.75	0	2.003%	84	0.75	0.25	0.75	1.976%					
40	0	0.75	0.25	2.003%	85	1	0.25	1	1.975%					
41	0.5	0.5	1	2.003%	86	0	0.5	1	1.975%					
42	0.25	0.5	0.25	2.003%	87	0.25	0.75	0	1.975%					
43	0.75	0.75	0.75	2.002%	88	0.75	0	1	1.974%					
44	0.25	0	1	2.001%	89	1	0.75	0.25	1.973%					
45	0.5	0	0	2.001%	90	0.25	0	0	1.973%					

Table D.31: F-measure of the FreqMax model on a monthly basis, for the last reply topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.5	0	1.520%	46	0.25	0.25	0.5	1.466%	91	0	0.5	0.75	1.446%
2	0.75	0.25	0.75	1.501%	47	0.5	0.75	0.75	1.466%	92	0.5	0.75	0.25	1.445%
3	1	1	0.5	1.496%	48	0.25	0.75	0.25	1.466%	93	0.5	0	0.5	1.445%
4	0.25	0.5	0.25	1.495%	49	0	0	0.25	1.466%	94	0.75	0.75	1	1.443%
5	1	0	0.25	1.490%	50	0.75	0.5	0.25	1.466%	95	0.25	0.5	0	1.443%
6	0	0.25	0	1.488%	51	0	1	0	1.465%	96	0	1	0.75	1.443%
7	0.25	1	1	1.488%	52	1	0	1	1.465%	97	1	0.25	0.5	1.443%
8	0	1	1	1.488%	53	0.5	1	0.5	1.465%	98	0.25	0	1	1.443%
9	0.25	0.5	0.75	1.487%	54	0.75	0.75	0.75	1.464%	99	0.5	0.25	0.75	1.442%
10	0.75	0	0.5	1.483%	55	0.75	1	0.5	1.463%	100	1	0.5	1	1.442%
11	1	0.25	1	1.483%	56	0.25	0.25	0.25	1.462%	101	0.5	0.25	0.5	1.441%
12	0.25	0.25	1	1.481%	57	0.75	1	0	1.462%	102	0.5	1	1	1.441%
13	0.5	0.25	1	1.480%	58	0.5	0.25	0.25	1.462%	103	1	0.5	0	1.440%
14	0.25	1	0.5	1.478%	59	0	0.75	0.75	1.461%	104	0.25	0.75	0.75	1.440%
15	0	0.25	0.25	1.478%	60	0.75	0.5	1	1.461%	105	0.75	0.75	0.25	1.440%
16	1	0.75	0	1.476%	61	0.75	0.25	1	1.460%	106	0.25	0	0.25	1.440%
17	0.25	0.75	1	1.476%	62	1	0.5	0.75	1.459%	107	0	0.75	0.5	1.439%
18	0.5	0.5	0.25	1.476%	63	0.5	0	0.75	1.459%	108	1	0.25	0.25	1.438%
19	0.25	0.5	1	1.476%	64	0	0	0.5	1.459%	109	0.75	0.25	0.5	1.438%
20	0.25	0	0	1.475%	65	1	0.5	0.5	1.458%	110	0.5	0	1	1.437%
21	0.75	1	0.25	1.475%	66	0.5	0.5	0.75	1.458%	111	0.5	0.25	0	1.437%
22	0	0.5	0.25	1.474%	67	1	0.75	0.5	1.457%	112	0.5	0.75	1	1.435%
23	0	0	0	1.474%	68	0	0.25	0.5	1.457%	113	0.75	1	0.75	1.435%
24	0	0.5	1	1.474%	69	0.75	0.75	0.5	1.456%	114	0.5	0.5	1	1.435%
25	0.5	0.75	0	1.473%	70	1	0	0	1.455%	115	0.75	0.5	0.5	1.434%
26	0	0.75	0.25	1.473%	71	0	1	0.5	1.455%	116	0	0.25	1	1.434%
27	0.25	1	0.25	1.472%	72	1	0	0.5	1.454%	117	0	0	0.75	1.434%
28	0.75	0	0.25	1.472%	73	0	0.75	0	1.454%	118	1	0.25	0	1.432%
29	0.5	0	0	1.472%	74	0.25	0.25	0.75	1.453%	119	1	1	0	1.431%
30	0	0.25	0.75	1.471%	75	1	0.75	1	1.453%	120	0.5	1	0.75	1.430%
31	1	1	0.75	1.471%	76	0.5	0	0.25	1.453%	121	0.75	0.5	0.75	1.428%
32	0.75	0.25	0.25	1.470%	77	0.25	0	0.75	1.453%	122	1	0.25	0.75	1.428%
33	0.25	0.75	0	1.470%	78	0.75	0.5	0	1.452%	123	0	1	0.25	1.423%
34	0.25	0.5	0.5	1.470%	79	0.75	0.75	0	1.452%	124	0.5	0.5	0	1.422%
35	0	0	1	1.470%	80	1	1	0.25	1.452%	125	0.25	1	0.75	1.420%
36	0.5	1	0	1.470%	81	1	0.75	0.75	1.452%					
37	0.25	0.25	0	1.469%	82	0	0.75	1	1.451%					
38	0.75	0	0.75	1.468%	83	0.5	1	0.25	1.451%					
39	0.25	0.75	0.5	1.468%	84	0.75	0.25	0	1.451%					
40	1	0	0.75	1.468%	85	0.75	0	1	1.450%					
41	1	1	1	1.468%	86	0.75	1	1	1.449%					
42	0.25	0	0.5	1.467%	87	1	0.5	0.25	1.449%					
43	0	0.5	0.5	1.467%	88	1	0.75	0.25	1.449%					
44	0.5	0.75	0.5	1.466%	89	0.75	0	0	1.448%					
45	0.25	1	0	1.466%	90	0.5	0.5	0.5	1.446%					

Table D.32: F-measure of the FreqLogit model on a monthly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	1	0.25	1	1.506%	46	0.25	0.5	0	1.470%	91	0.25	1	0	1.450%
2	0.75	1	0.75	1.506%	47	0.25	0.5	0.75	1.470%	92	1	0.5	0	1.449%
3	0.5	0.5	0.25	1.504%	48	0.5	0	0.5	1.470%	93	0.5	1	0.5	1.448%
4	0.25	0.5	0.5	1.502%	49	0	0.25	0.25	1.469%	94	0.5	0.25	0.25	1.448%
5	0	0.75	0.5	1.498%	50	0.25	0.75	0.5	1.469%	95	0.75	0.25	0.25	1.448%
6	0.75	0	0	1.492%	51	0.75	0.25	1	1.468%	96	0.5	0.5	0.75	1.448%
7	1	0.75	0.25	1.492%	52	0.25	0	0.5	1.468%	97	0.5	1	0.75	1.447%
8	0.25	0.75	0.75	1.490%	53	0.75	0.5	0	1.467%	98	1	0.75	0	1.447%
9	0.75	0.75	0.75	1.489%	54	0.5	1	1	1.466%	99	1	0.5	0.75	1.447%
10	1	0	0.75	1.488%	55	0.75	0.75	1	1.466%	100	0.25	1	1	1.447%
11	0.75	0.75	0.5	1.488%	56	0.25	1	0.25	1.465%	101	0	0	0	1.447%
12	1	0.5	0.5	1.487%	57	0.5	0.75	0.5	1.465%	102	1	1	1	1.446%
13	0.75	0.75	0	1.487%	58	0	1	0.5	1.465%	103	0	1	0.25	1.445%
14	0	1	0.75	1.486%	59	0.25	0.25	0.5	1.465%	104	0.25	0.75	1	1.444%
15	0	0.5	1	1.486%	60	1	0.5	1	1.465%	105	0	0	0.25	1.444%
16	0	0.5	0	1.485%	61	0.75	0.5	0.75	1.464%	106	0.75	0	0.75	1.444%
17	0.75	0.5	0.25	1.484%	62	0.25	1	0.5	1.464%	107	1	0.75	1	1.443%
18	0.25	0.25	0.25	1.483%	63	0	0.5	0.5	1.463%	108	0.5	0.25	1	1.441%
19	0.75	1	0	1.482%	64	1	0.75	0.75	1.463%	109	1	0	1	1.441%
20	0.5	1	0.25	1.481%	65	0.25	0.5	1	1.463%	110	0.5	0.75	0	1.440%
21	0.5	0.25	0	1.480%	66	0.5	0.25	0.75	1.463%	111	0.5	0.75	1	1.439%
22	0.75	0.5	0.5	1.480%	67	1	0	0.5	1.463%	112	1	0	0.25	1.435%
23	0.75	1	1	1.479%	68	1	0.25	0.5	1.462%	113	1	0.5	0.25	1.434%
24	0.5	0.25	0.5	1.479%	69	0.25	0.25	0.75	1.462%	114	1	0.25	0	1.434%
25	0	0.25	1	1.479%	70	0.5	0.75	0.75	1.461%	115	0	0.5	0.75	1.433%
26	0.75	0	0.5	1.478%	71	0.75	0.5	1	1.460%	116	1	1	0.75	1.433%
27	0	0.75	1	1.476%	72	1	0.75	0.5	1.460%	117	0	0	0.5	1.433%
28	0.75	1	0.25	1.475%	73	0	0	1	1.459%	118	0.25	0.25	0	1.433%
29	0.75	0.75	0.25	1.475%	74	0.25	0	0.75	1.459%	119	0.25	0.75	0.25	1.432%
30	0.75	0.25	0.5	1.475%	75	0	0.75	0.75	1.459%	120	0	0.75	0.25	1.432%
31	0.25	0.75	0	1.474%	76	0	0.5	0.25	1.459%	121	0	0.25	0.5	1.432%
32	0.5	0.5	0	1.474%	77	0.75	0.25	0.75	1.458%	122	0.25	0	0.25	1.425%
33	0.25	0	1	1.474%	78	0.75	0	1	1.457%	123	0.5	0	0.25	1.423%
34	0.5	0	0.75	1.474%	79	0	0.25	0.75	1.456%	124	1	1	0.5	1.415%
35	0.5	0.75	0.25	1.473%	80	0.25	1	0.75	1.455%	125	0	1	1	1.415%
36	0.75	0.25	0	1.473%	81	0	0	0.75	1.455%					
37	0.5	0.5	1	1.473%	82	0.75	0	0.25	1.455%					
38	0.25	0.25	1	1.472%	83	1	1	0.25	1.455%					
39	1	0.25	0.25	1.472%	84	1	1	0	1.455%					
40	1	0	0	1.471%	85	0	0.75	0	1.454%					
41	0.75	1	0.5	1.471%	86	0.5	0.5	0.5	1.453%					
42	0	1	0	1.471%	87	0	0.25	0	1.453%					
43	0.5	0	0	1.471%	88	0.5	0	1	1.453%					
44	0.25	0	0	1.471%	89	0.5	1	0	1.452%					
45	0.25	0.5	0.25	1.471%	90	1	0.25	0.75	1.451%					

Table D.33: F-measure of the FreqRandom model on a monthly basis, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.25	0.75	0	7.341%	46	0.25	0.75	0.5	7.247%	91	1	0	0.75	7.200%
2	0.75	0.5	0.5	7.337%	47	0	1	1	7.246%	92	0.25	1	0.5	7.198%
3	0	1	0.25	7.323%	48	0	0.25	0	7.245%	93	1	0	0.25	7.197%
4	0.25	0.5	0.25	7.317%	49	0.75	0.25	0.75	7.244%	94	1	1	0.5	7.195%
5	0.5	0.75	0.25	7.309%	50	0.5	0.75	0.5	7.244%	95	1	0.75	0.5	7.194%
6	0	0.25	0.25	7.306%	51	0.25	0	0.75	7.244%	96	0.25	0.75	0.75	7.194%
7	0.75	1	0	7.301%	52	1	0.25	0.5	7.244%	97	1	0	1	7.193%
8	0.75	0.5	0.25	7.295%	53	0.25	0.25	0	7.243%	98	0.75	1	1	7.192%
9	0.75	0.5	1	7.292%	54	0.75	0.75	0.75	7.242%	99	0.25	0.5	0.75	7.190%
10	0.75	0.75	1	7.288%	55	1	0.75	1	7.241%	100	0.75	0.75	0.25	7.189%
11	0.75	0.5	0	7.286%	56	1	1	0.75	7.241%	101	1	0.75	0	7.188%
12	0	0.75	1	7.286%	57	1	0.75	0.25	7.239%	102	1	1	0.25	7.187%
13	0.25	1	0	7.283%	58	0.25	1	1	7.238%	103	0.5	1	1	7.184%
14	0.75	0.75	0	7.281%	59	0.75	0.25	0.5	7.238%	104	0.25	0.5	1	7.184%
15	0.75	0	0.5	7.280%	60	0	0.25	0.75	7.238%	105	0.5	0.25	0.25	7.184%
16	0.25	0.25	1	7.280%	61	0.25	0	0.5	7.238%	106	0.5	0.5	1	7.182%
17	0	0.5	0	7.280%	62	0.75	1	0.25	7.237%	107	1	0.25	0.75	7.180%
18	0.5	0.5	0.75	7.279%	63	1	0.75	0.75	7.237%	108	0	0	0.75	7.179%
19	0.75	0.25	0.25	7.279%	64	0.75	0.75	0.5	7.236%	109	1	0.5	0	7.179%
20	0	1	0	7.278%	65	0.5	1	0.25	7.236%	110	0	0	0	7.178%
21	0.25	0.25	0.25	7.273%	66	0	0.25	0.5	7.235%	111	1	0.5	0.75	7.173%
22	1	0.5	0.5	7.272%	67	0.5	0.75	0.75	7.233%	112	0.75	0.25	1	7.173%
23	1	1	0	7.270%	68	0.75	0	0	7.232%	113	0	0.5	0.5	7.169%
24	0.5	1	0.5	7.269%	69	0.25	0.25	0.5	7.229%	114	0	0.75	0.75	7.168%
25	0.5	0.5	0.5	7.268%	70	0.25	0.5	0.5	7.229%	115	0.5	0.75	0	7.166%
26	0.5	0	0	7.266%	71	0.75	0.5	0.75	7.227%	116	0.75	0.25	0	7.161%
27	1	0.25	0	7.265%	72	0.5	0.75	1	7.227%	117	0.25	1	0.25	7.160%
28	0	0	0.25	7.263%	73	0.5	0.25	0.75	7.224%	118	0.5	0.25	0	7.155%
29	0.5	1	0.75	7.262%	74	0	0.5	0.75	7.222%	119	0	0	0.5	7.151%
30	1	0.25	1	7.262%	75	0.25	0.25	0.75	7.221%	120	0.25	0	0	7.142%
31	0	0.75	0.5	7.260%	76	0.25	0	0.25	7.221%	121	0.75	1	0.75	7.139%
32	0.5	0	0.25	7.259%	77	0	0.75	0	7.219%	122	0.5	0.25	0.5	7.138%
33	0	0	1	7.259%	78	0.25	0.75	1	7.219%	123	0.75	0	1	7.137%
34	0.5	0	1	7.258%	79	0	0.75	0.25	7.219%	124	1	0.25	0.25	7.131%
35	0.5	0	0.5	7.258%	80	0.75	1	0.5	7.217%	125	0.5	0.5	0	7.109%
36	1	1	1	7.255%	81	0.5	0.25	1	7.214%					
37	0.25	0	1	7.254%	82	0.5	0	0.75	7.213%					
38	0	1	0.5	7.252%	83	1	0	0	7.213%					
39	1	0.5	0.25	7.251%	84	0.5	0.5	0.25	7.213%					
40	0.75	0	0.25	7.249%	85	0.25	0.5	0	7.213%					
41	0	0.5	1	7.248%	86	0.75	0	0.75	7.209%					
42	0	0.5	0.25	7.248%	87	1	0.5	1	7.209%					
43	0	1	0.75	7.248%	88	1	0	0.5	7.209%					
44	0	0.25	1	7.247%	89	0.25	0.75	0.25	7.208%					
45	0.5	1	0	7.247%	90	0.25	1	0.75	7.202%					

Table D.34: F-measure of the FreqLCA model over the whole period of simulation, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0.25	0	8.087%	46	0.5	0.5	0	7.960%	91	0.5	0	0.25	7.912%
2	1	0.25	0	8.080%	47	0.5	0	0.75	7.959%	92	0.25	0.75	0.75	7.912%
3	0.5	0.25	0.5	8.076%	48	0	0	0.5	7.959%	93	0	0.25	0.5	7.911%
4	0.5	0.5	0.25	8.056%	49	0.5	0.75	1	7.957%	94	0.75	0.5	0.75	7.910%
5	0	0.5	0	8.044%	50	0.25	0.25	1	7.955%	95	0	0.75	0.25	7.908%
6	0.5	1	0.5	8.040%	51	0.25	0	0	7.954%	96	0.25	0.25	0.5	7.908%
7	0.5	0	0	8.035%	52	1	0.75	0.5	7.952%	97	1	0.25	0.5	7.907%
8	1	0	0.75	8.031%	53	0.25	0	0.75	7.952%	98	1	0	0.25	7.906%
9	1	0	1	8.023%	54	0.75	0.75	0	7.950%	99	0	0	0.75	7.904%
10	0	0.5	0.25	8.023%	55	0.75	1	0.5	7.950%	100	0.25	0.75	1	7.904%
11	0.75	0.25	0.25	8.021%	56	1	0.5	0.5	7.949%	101	0.75	1	0	7.903%
12	0.5	0.75	0.75	8.020%	57	1	0.25	0.25	7.948%	102	1	0.75	1	7.903%
13	0.25	0.25	0	8.017%	58	0.25	0	0.25	7.948%	103	0	0.5	0.75	7.903%
14	0.75	0.25	0	8.016%	59	0	0.25	0.25	7.948%	104	1	0	0	7.900%
15	0.75	0.5	0.25	8.015%	60	0	1	0.25	7.947%	105	1	1	0.5	7.895%
16	0.5	1	0	8.015%	61	0	0	0	7.947%	106	0.75	0	1	7.892%
17	0.75	0	0.75	8.015%	62	0.5	0.75	0	7.947%	107	0.75	0	0.5	7.890%
18	0.5	0.25	0.25	8.013%	63	0	0.75	1	7.945%	108	0.25	1	0.5	7.889%
19	0	0.75	0.75	8.012%	64	0.5	0.75	0.25	7.944%	109	0.5	0.25	0	7.888%
20	0.25	0.5	0.25	8.012%	65	1	0.25	0.75	7.943%	110	0	0.25	1	7.885%
21	0.75	0.25	0.75	8.005%	66	1	0.75	0.75	7.942%	111	0	0.5	0.5	7.883%
22	1	0	0.5	8.004%	67	0	0	0.25	7.938%	112	0.25	0.75	0	7.874%
23	1	0.75	0	8.003%	68	0.25	0.25	0.75	7.937%	113	1	0.5	0.75	7.869%
24	0	0.75	0	8.003%	69	0.25	0.75	0.5	7.936%	114	0.75	0.25	0.5	7.869%
25	0.25	0.5	0	8.001%	70	0.5	1	0.25	7.936%	115	0	0.5	1	7.868%
26	0.25	1	0.75	7.998%	71	0	0	1	7.935%	116	0.5	0.5	0.75	7.864%
27	0.75	0	0.25	7.996%	72	0.5	0	1	7.934%	117	0.25	0.5	0.5	7.863%
28	0.25	0.75	0.25	7.995%	73	0.75	1	0.25	7.933%	118	0.5	1	0.75	7.843%
29	0.5	0	0.5	7.995%	74	0.5	0.5	1	7.933%	119	0.75	0.25	1	7.841%
30	0.5	0.75	0.5	7.993%	75	0.75	0.5	0.5	7.931%	120	0.5	1	1	7.840%
31	0	0.75	0.5	7.990%	76	1	0.5	0	7.931%	121	0.5	0.25	1	7.826%
32	1	1	0.75	7.990%	77	1	0.25	1	7.930%	122	1	1	1	7.794%
33	0.25	0	0.5	7.988%	78	0.75	0.75	0.75	7.929%	123	1	0.5	1	7.792%
34	0	1	0.75	7.983%	79	0	1	1	7.926%	124	0.75	0.5	1	7.785%
35	0.25	0.5	0.75	7.979%	80	0.75	1	0.75	7.924%	125	0.25	0.5	1	7.779%
36	0.75	0.75	1	7.977%	81	0.75	0	0	7.923%					
37	0.75	0.75	0.25	7.976%	82	1	1	0	7.921%					
38	0.75	0.5	0	7.976%	83	0.75	1	1	7.920%					
39	0	0.25	0.75	7.976%	84	1	1	0.25	7.920%					
40	0.25	1	1	7.974%	85	0.5	0.5	0.5	7.920%					
41	0.25	1	0.25	7.971%	86	0	1	0.5	7.917%					
42	1	0.75	0.25	7.969%	87	0.75	0.75	0.5	7.914%					
43	0.25	0	1	7.966%	88	0	1	0	7.914%					
44	1	0.5	0.25	7.963%	89	0.25	1	0	7.913%					
45	0.25	0.25	0.25	7.963%	90	0.5	0.25	0.75	7.913%					

Table D.35: F-measure of the FreqMax model over the whole period of simulation, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0	0	0.25	7.418%	46	0.25	0	0	7.338%	91	0	0.5	1	7.290%
2	0.75	0.25	0.75	7.415%	47	0	0.25	0	7.338%	92	0.75	0.75	0.75	7.290%
3	0.25	1	0.25	7.413%	48	0.25	0.75	0.5	7.337%	93	0.5	1	0.75	7.288%
4	0.75	0.5	1	7.412%	49	0	1	0	7.336%	94	0.75	0.75	0	7.288%
5	0.5	1	0.5	7.403%	50	1	1	1	7.335%	95	0.75	0	1	7.287%
6	0	0.25	0.25	7.402%	51	0.25	0.75	0.25	7.334%	96	0.75	0.25	0	7.285%
7	0.25	0.75	0	7.394%	52	0.5	0.5	0.75	7.333%	97	1	0	0.25	7.284%
8	1	1	0.75	7.392%	53	0.75	1	0.75	7.332%	98	0.5	0	0.75	7.283%
9	0.5	0.5	0.25	7.391%	54	0.5	0.75	0.5	7.330%	99	0	0	0.5	7.283%
10	0	0.25	0.75	7.387%	55	1	0	0.5	7.329%	100	0.5	0.75	1	7.281%
11	1	1	0.5	7.385%	56	0.5	0.25	0.25	7.329%	101	0	0.75	0.75	7.280%
12	0.25	0.5	0.25	7.384%	57	0.25	0.75	1	7.326%	102	0	0.75	1	7.276%
13	0.5	1	1	7.384%	58	0.25	0.25	0.25	7.326%	103	0.5	0.75	0.25	7.274%
14	0.25	1	0	7.383%	59	0.25	0	0.75	7.323%	104	0	1	0.5	7.273%
15	0.25	1	0.5	7.383%	60	1	0.5	0.25	7.322%	105	1	1	0	7.272%
16	1	0.25	1	7.381%	61	0	1	0.25	7.322%	106	0.25	0	0.5	7.272%
17	0.25	1	0.75	7.380%	62	0	1	1	7.322%	107	1	0.25	0.25	7.270%
18	0.75	0	0.25	7.379%	63	0.5	0.75	0	7.321%	108	0.5	0.25	0	7.268%
19	0.75	0.75	1	7.379%	64	0.75	0.25	0.25	7.321%	109	0.25	0	1	7.267%
20	1	0.75	0	7.377%	65	0.75	0.25	0.5	7.319%	110	0.5	0.5	0.5	7.265%
21	1	0	0.75	7.377%	66	0	0.75	0.25	7.318%	111	0.5	0	0.25	7.264%
22	0.75	0.75	0.5	7.377%	67	0.5	0.25	0.75	7.317%	112	1	0.25	0.75	7.260%
23	0.75	0	0.5	7.377%	68	0	0.75	0	7.315%	113	0.5	1	0.25	7.257%
24	0.75	1	0.5	7.376%	69	0	0.75	0.5	7.313%	114	1	0.25	0.5	7.257%
25	1	0.75	0.25	7.375%	70	0.75	0.5	0.75	7.313%	115	0.25	0.25	0.75	7.257%
26	0.25	0.5	1	7.368%	71	0.25	0.75	0.75	7.313%	116	1	0.5	1	7.249%
27	0.75	1	0.25	7.368%	72	0.75	0.5	0.25	7.312%	117	0	0.25	0.5	7.248%
28	0.75	0	0.75	7.367%	73	0.25	1	1	7.311%	118	0.75	0.5	0	7.246%
29	0	0.5	0.25	7.367%	74	0.5	1	0	7.308%	119	1	0.25	0	7.241%
30	0.5	0	0	7.362%	75	1	1	0.25	7.308%	120	0.75	0	0	7.240%
31	1	0.75	1	7.360%	76	0.25	0.25	1	7.305%	121	0.75	1	1	7.228%
32	0	0.5	0.75	7.360%	77	0.25	0.25	0.5	7.304%	122	0.5	0.25	0.5	7.221%
33	0.25	0.25	0	7.359%	78	0	0	0.75	7.303%	123	0.75	0.25	1	7.215%
34	0	0.5	0	7.359%	79	1	0.75	0.5	7.302%	124	0.75	0.5	0.5	7.203%
35	0.25	0.5	0	7.359%	80	1	0.75	0.75	7.301%	125	0.5	0.5	1	7.201%
36	0.25	0.5	0.75	7.358%	81	0.5	0.25	1	7.300%					
37	1	0	1	7.357%	82	0	1	0.75	7.299%					
38	0	0	1	7.357%	83	0.75	0.75	0.25	7.298%					
39	1	0.5	0.75	7.351%	84	0.5	0.5	0	7.295%					
40	0	0	0	7.349%	85	0.25	0	0.25	7.295%					
41	1	0	0	7.347%	86	1	0.5	0.5	7.295%					
42	0.25	0.5	0.5	7.347%	87	0.5	0	0.5	7.294%					
43	0.75	1	0	7.346%	88	0.5	0.75	0.75	7.294%					
44	0	0.5	0.5	7.344%	89	0.5	0	1	7.293%					
45	0	0.25	1	7.339%	90	1	0.5	0	7.291%					

Table D.36: F-measure of the FreqLogit model over the whole period of simulation, for the last post topology.

Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}	Rank	c_π	c_μ	c_ν	F_{meas}
1	0.75	0.5	0.25	7.447%	46	0.75	0.25	0.75	7.355%	91	1	0.25	0.5	7.305%
2	0	0.5	0	7.434%	47	0.25	1	0.25	7.355%	92	0.25	0.75	0.25	7.305%
3	0.25	0.75	0	7.430%	48	1	0.25	1	7.355%	93	0	0.25	0.25	7.304%
4	1	0	0.5	7.426%	49	0.5	0.5	0	7.354%	94	1	0.25	0	7.304%
5	1	0.5	0.5	7.422%	50	0.75	0.75	1	7.351%	95	0.75	0.25	0	7.304%
6	0.75	1	0	7.411%	51	0	0.25	0.75	7.351%	96	0.25	0.75	1	7.304%
7	0.25	0	0.75	7.410%	52	1	0	0	7.349%	97	0.5	0	1	7.302%
8	0.75	0.25	1	7.410%	53	0.25	0.5	0.75	7.348%	98	1	0	0.25	7.300%
9	0.75	0	0.5	7.405%	54	0.75	0	0.25	7.345%	99	0.25	0.25	0.5	7.299%
10	0.5	0.75	0.5	7.402%	55	0.25	1	0	7.345%	100	0.5	0.75	0.25	7.298%
11	0.25	0.25	1	7.400%	56	0.25	1	0.5	7.343%	101	0.5	1	0	7.298%
12	0.25	0.75	0.75	7.398%	57	0.75	0	1	7.342%	102	1	0.5	0.75	7.297%
13	0	0.75	1	7.396%	58	0.75	0	0.75	7.341%	103	1	1	0.75	7.294%
14	0.75	0.75	0.75	7.395%	59	0.5	0.75	0	7.339%	104	0.25	0.5	1	7.292%
15	0.75	0.75	0.5	7.393%	60	0.75	0.5	0	7.338%	105	0.5	0	0.5	7.292%
16	0.75	1	0.75	7.392%	61	0.5	0.5	1	7.338%	106	0	0.75	0.75	7.288%
17	1	0	0.75	7.391%	62	0.75	0.75	0.25	7.337%	107	0.5	0.25	0.25	7.288%
18	0.25	1	0.75	7.388%	63	1	0.75	0	7.337%	108	0.5	0.25	0.75	7.286%
19	0.75	1	1	7.385%	64	0	0	1	7.337%	109	0.5	1	0.75	7.283%
20	0.5	1	0.25	7.381%	65	0.5	0.75	1	7.337%	110	1	1	1	7.283%
21	0.75	1	0.5	7.381%	66	0.25	0.5	0.5	7.336%	111	0.5	0	0.25	7.281%
22	0.75	0.75	0	7.380%	67	0.25	0	0	7.336%	112	1	1	0.5	7.278%
23	0.5	0.5	0.75	7.380%	68	1	0.75	0.75	7.336%	113	0.25	0.25	0	7.276%
24	0	1	0.75	7.379%	69	1	0.25	0.75	7.335%	114	0.25	0	0.25	7.274%
25	0.75	0.5	0.75	7.376%	70	0	0.5	1	7.335%	115	0.75	0.25	0.25	7.272%
26	0.25	0	1	7.374%	71	0.5	0.75	0.75	7.332%	116	1	1	0.25	7.271%
27	0.25	0	0.5	7.374%	72	1	1	0	7.331%	117	0	0.5	0.75	7.270%
28	0.25	0.75	0.5	7.374%	73	0.25	0.25	0.75	7.331%	118	0	0.5	0.25	7.268%
29	0	0	0	7.374%	74	0.75	0.5	1	7.328%	119	1	0.25	0.25	7.268%
30	1	0.75	0.25	7.371%	75	0.5	0.5	0.25	7.328%	120	1	0.75	0.5	7.266%
31	0.25	1	1	7.370%	76	0.5	0	0	7.327%	121	0	0.25	0	7.257%
32	0.75	1	0.25	7.369%	77	0	0	0.5	7.323%	122	0.5	1	0.5	7.256%
33	0.75	0	0	7.369%	78	0.5	1	1	7.322%	123	0	1	0.25	7.250%
34	1	0.5	1	7.365%	79	0.5	0.25	1	7.321%	124	0	0.25	0.5	7.238%
35	0	0	0.75	7.361%	80	0	0.75	0	7.317%	125	0	1	1	7.219%
36	0	0.25	1	7.361%	81	0.25	0.25	0.25	7.315%					
37	1	0.5	0	7.360%	82	0.5	0.5	0.5	7.315%					
38	1	0.75	1	7.359%	83	0	1	0.5	7.312%					
39	0.75	0.25	0.5	7.359%	84	0.25	0.5	0.25	7.312%					
40	0.5	0.25	0	7.359%	85	0	0	0.25	7.311%					
41	0.5	0.25	0.5	7.358%	86	1	0	1	7.309%					
42	0	0.5	0.5	7.357%	87	0	0.75	0.25	7.309%					
43	0	0.75	0.5	7.357%	88	0.25	0.5	0	7.308%					
44	0.75	0.5	0.5	7.357%	89	0.5	0	0.75	7.307%					
45	0	1	0	7.356%	90	1	0.5	0.25	7.307%					

Table D.37: F-measure of the FreqRandom model over the whole period of simulation, for the last post topology.

Appendix E

Contents variance

Table E.1: Trend results index.

c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$
0	0	0	-7.91E-007	0	1	0.25	7.77E-007	0	0.75	0.75	-1.44E-006
0.25	0	0	3.99E-007	0.25	1	0.25	9.11E-008	0.25	0.75	0.75	-5.51E-007
0.5	0	0	-7.78E-007	0.5	1	0.25	-1.09E-006	0.5	0.75	0.75	-1.28E-006
0.75	0	0	-1.45E-006	0.75	1	0.25	1.10E-006	0.75	0.75	0.75	-6.42E-007
1	0	0	4.09E-007	1	1	0.25	1.17E-006	1	0.75	0.75	-4.03E-007
0	0.25	0	-5.77E-006	0	0	0.5	2.70E-006	0	1	0.75	4.34E-007
0.25	0.25	0	-5.36E-006	0.25	0	0.5	1.31E-006	0.25	1	0.75	1.20E-006
0.5	0.25	0	-1.69E-006	0.5	0	0.5	-9.10E-007	0.5	1	0.75	2.50E-006
0.75	0.25	0	-1.60E-006	0.75	0	0.5	-2.14E-006	0.75	1	0.75	2.44E-007
1	0.25	0	7.96E-007	1	0	0.5	3.39E-007	1	1	0.75	2.28E-006
0	0.5	0	-1.65E-006	0	0.25	0.5	-5.54E-006	0	0	1	3.34E-006
0.25	0.5	0	-1.31E-006	0.25	0.25	0.5	-4.44E-006	0.25	0	1	1.10E-006
0.5	0.5	0	-1.06E-006	0.5	0.25	0.5	-3.59E-006	0.5	0	1	-1.76E-006
0.75	0.5	0	-1.66E-006	0.75	0.25	0.5	-6.82E-007	0.75	0	1	-2.52E-006
1	0.5	0	1.67E-006	1	0.25	0.5	-2.49E-007	1	0	1	-3.23E-007
0	0.75	0	-9.69E-007	0	0.5	0.5	-3.43E-006	0	0.25	1	-5.08E-006
0.25	0.75	0	-1.96E-006	0.25	0.5	0.5	-8.62E-007	0.25	0.25	1	-4.17E-006
0.5	0.75	0	1.52E-006	0.5	0.5	0.5	-2.38E-006	0.5	0.25	1	-3.41E-006
0.75	0.75	0	1.21E-006	0.75	0.5	0.5	1.43E-006	0.75	0.25	1	-3.50E-007
1	0.75	0	4.24E-008	1	0.5	0.5	2.06E-006	1	0.25	1	1.02E-006
0	1	0	1.51E-006	0	0.75	0.5	-1.04E-006	0	0.5	1	-9.69E-007
0.25	1	0	5.55E-008	0.25	0.75	0.5	-3.11E-006	0.25	0.5	1	-2.11E-006
0.5	1	0	9.30E-007	0.5	0.75	0.5	1.37E-006	0.5	0.5	1	-1.16E-006
0.75	1	0	-5.34E-007	0.75	0.75	0.5	4.00E-007	0.75	0.5	1	-2.32E-006
1	1	0	-3.55E-007	1	0.75	0.5	5.08E-007	1	0.5	1	4.47E-007
0	0	0.25	-2.94E-007	0	1	0.5	7.32E-007	0	0.75	1	-4.06E-007
0.25	0	0.25	2.44E-006	0.25	1	0.5	5.95E-007	0.25	0.75	1	2.30E-007
0.5	0	0.25	-1.68E-006	0.5	1	0.5	3.23E-006	0.5	0.75	1	1.35E-006
0.75	0	0.25	-2.99E-006	0.75	1	0.5	3.53E-006	0.75	0.75	1	2.01E-006
1	0	0.25	4.00E-007	1	1	0.5	3.67E-006	1	0.75	1	2.11E-006
0	0.25	0.25	-5.60E-006	0	0	0.75	2.40E-006	0	1	1	2.47E-006
0.25	0.25	0.25	-4.57E-006	0.25	0	0.75	2.25E-006	0.25	1	1	-5.06E-007
0.5	0.25	0.25	-3.50E-006	0.5	0	0.75	-1.19E-006	0.5	1	1	1.22E-006
0.75	0.25	0.25	-1.50E-006	0.75	0	0.75	-1.47E-006	0.75	1	1	1.15E-006
1	0.25	0.25	4.29E-007	1	0	0.75	9.09E-007	1	1	1	-3.94E-008
0	0.5	0.25	-2.72E-006	0	0.25	0.75	-5.31E-006				
0.25	0.5	0.25	-2.27E-006	0.25	0.25	0.75	-2.23E-006				
0.5	0.5	0.25	-6.60E-007	0.5	0.25	0.75	-2.40E-006				
0.75	0.5	0.25	2.11E-007	0.75	0.25	0.75	-2.82E-006				
1	0.5	0.25	1.10E-006	1	0.25	0.75	1.07E-006				
0	0.75	0.25	-1.79E-006	0	0.5	0.75	-3.85E-006				
0.25	0.75	0.25	-1.03E-007	0.25	0.5	0.75	-8.66E-007				
0.5	0.75	0.25	-1.35E-006	0.5	0.5	0.75	-2.10E-006				
0.75	0.75	0.25	4.40E-007	0.75	0.5	0.75	-2.07E-006				
1	0.75	0.25	2.08E-006	1	0.5	0.75	5.37E-007				

Table E.2: Trend on a weekly basis of the average topic weights variance for the FreqLCA model.

c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$
0	0	0	2.77E-006	0	1	0.25	1.93E-006	0	0.75	0.75	1.34E-006
0.25	0	0	4.75E-006	0.25	1	0.25	1.00E-006	0.25	0.75	0.75	8.32E-007
0.5	0	0	4.50E-006	0.5	1	0.25	1.87E-006	0.5	0.75	0.75	1.06E-006
0.75	0	0	1.46E-005	0.75	1	0.25	1.59E-006	0.75	0.75	0.75	2.89E-007
1	0	0	3.54E-006	1	1	0.25	8.47E-007	1	0.75	0.75	1.98E-006
0	0.25	0	3.46E-006	0	0	0.5	2.33E-006	0	1	0.75	1.81E-006
0.25	0.25	0	2.18E-006	0.25	0	0.5	5.77E-006	0.25	1	0.75	1.86E-006
0.5	0.25	0	2.21E-006	0.5	0	0.5	8.58E-006	0.5	1	0.75	2.62E-006
0.75	0.25	0	1.67E-006	0.75	0	0.5	1.35E-005	0.75	1	0.75	1.53E-006
1	0.25	0	2.87E-006	1	0	0.5	-1.02E-006	1	1	0.75	2.10E-006
0	0.5	0	2.05E-006	0	0.25	0.5	1.30E-006	0	0	1	2.37E-006
0.25	0.5	0	1.60E-006	0.25	0.25	0.5	1.27E-006	0.25	0	1	3.60E-006
0.5	0.5	0	2.16E-006	0.5	0.25	0.5	2.48E-006	0.5	0	1	9.77E-006
0.75	0.5	0	1.33E-006	0.75	0.25	0.5	8.97E-007	0.75	0	1	1.36E-005
1	0.5	0	1.49E-006	1	0.25	0.5	2.47E-006	1	0	1	9.92E-007
0	0.75	0	1.59E-006	0	0.5	0.5	2.98E-006	0	0.25	1	3.03E-006
0.25	0.75	0	3.09E-006	0.25	0.5	0.5	2.16E-006	0.25	0.25	1	6.07E-007
0.5	0.75	0	2.29E-006	0.5	0.5	0.5	2.56E-006	0.5	0.25	1	4.50E-007
0.75	0.75	0	1.22E-006	0.75	0.5	0.5	1.59E-006	0.75	0.25	1	1.55E-006
1	0.75	0	8.59E-007	1	0.5	0.5	2.80E-006	1	0.25	1	5.49E-007
0	1	0	1.72E-006	0	0.75	0.5	2.49E-006	0	0.5	1	2.40E-006
0.25	1	0	1.85E-006	0.25	0.75	0.5	3.94E-007	0.25	0.5	1	1.94E-006
0.5	1	0	1.40E-006	0.5	0.75	0.5	1.52E-006	0.5	0.5	1	2.83E-006
0.75	1	0	2.19E-006	0.75	0.75	0.5	-9.64E-007	0.75	0.5	1	1.46E-006
1	1	0	2.75E-007	1	0.75	0.5	3.21E-006	1	0.5	1	2.43E-006
0	0	0.25	1.29E-006	0	1	0.5	1.95E-006	0	0.75	1	2.41E-006
0.25	0	0.25	4.80E-006	0.25	1	0.5	6.13E-007	0.25	0.75	1	2.46E-006
0.5	0	0.25	7.03E-006	0.5	1	0.5	-9.99E-009	0.5	0.75	1	2.57E-006
0.75	0	0.25	1.09E-005	0.75	1	0.5	2.92E-006	0.75	0.75	1	2.93E-006
1	0	0.25	2.03E-006	1	1	0.5	2.89E-006	1	0.75	1	1.11E-006
0	0.25	0.25	1.16E-006	0	0	0.75	2.67E-006	0	1	1	1.36E-006
0.25	0.25	0.25	1.73E-006	0.25	0	0.75	4.57E-006	0.25	1	1	1.43E-007
0.5	0.25	0.25	2.26E-006	0.5	0	0.75	6.69E-006	0.5	1	1	3.26E-006
0.75	0.25	0.25	1.78E-006	0.75	0	0.75	1.42E-005	0.75	1	1	9.12E-007
1	0.25	0.25	1.25E-006	1	0	0.75	7.50E-007	1	1	1	3.12E-006
0	0.5	0.25	3.15E-006	0	0.25	0.75	1.93E-006				
0.25	0.5	0.25	2.74E-006	0.25	0.25	0.75	2.93E-006				
0.5	0.5	0.25	3.62E-006	0.5	0.25	0.75	2.54E-006				
0.75	0.5	0.25	1.37E-006	0.75	0.25	0.75	3.76E-006				
1	0.5	0.25	3.32E-006	1	0.25	0.75	3.31E-006				
0	0.75	0.25	1.13E-007	0	0.5	0.75	2.77E-006				
0.25	0.75	0.25	1.61E-006	0.25	0.5	0.75	2.44E-006				
0.5	0.75	0.25	1.22E-006	0.5	0.5	0.75	3.03E-006				
0.75	0.75	0.25	3.38E-006	0.75	0.5	0.75	1.72E-006				
1	0.75	0.25	1.37E-006	1	0.5	0.75	2.66E-006				

Table E.3: Trend on a weekly basis of the average topic weights variance for the FreqMax model.

c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$
0	0	0	1.20E-006	0	1	0.25	2.24E-006	0	0.75	0.75	-5.15E-007
0.25	0	0	1.58E-006	0.25	1	0.25	1.36E-006	0.25	0.75	0.75	4.71E-007
0.5	0	0	-1.72E-006	0.5	1	0.25	2.04E-006	0.5	0.75	0.75	2.09E-007
0.75	0	0	-1.87E-006	0.75	1	0.25	1.16E-006	0.75	0.75	0.75	1.90E-006
1	0	0	2.44E-006	1	1	0.25	3.25E-006	1	0.75	0.75	1.85E-006
0	0.25	0	-3.72E-006	0	0	0.5	8.29E-007	0	1	0.75	1.41E-006
0.25	0.25	0	-3.68E-006	0.25	0	0.5	-1.85E-007	0.25	1	0.75	2.84E-006
0.5	0.25	0	-1.15E-006	0.5	0	0.5	-2.86E-006	0.5	1	0.75	4.81E-006
0.75	0.25	0	-2.23E-007	0.75	0	0.5	-2.37E-006	0.75	1	0.75	2.82E-006
1	0.25	0	1.88E-006	1	0	0.5	1.11E-006	1	1	0.75	1.23E-006
0	0.5	0	-1.66E-006	0	0.25	0.5	-5.37E-006	0	0	1	8.98E-007
0.25	0.5	0	-7.78E-007	0.25	0.25	0.5	-3.49E-006	0.25	0	1	1.29E-006
0.5	0.5	0	-8.74E-007	0.5	0.25	0.5	-2.94E-006	0.5	0	1	-8.99E-007
0.75	0.5	0	8.27E-007	0.75	0.25	0.5	-6.32E-007	0.75	0	1	-1.18E-006
1	0.5	0	2.67E-006	1	0.25	0.5	3.15E-006	1	0	1	3.03E-006
0	0.75	0	5.40E-007	0	0.5	0.5	-1.86E-006	0	0.25	1	-4.83E-006
0.25	0.75	0	5.97E-007	0.25	0.5	0.5	-1.54E-006	0.25	0.25	1	-2.82E-006
0.5	0.75	0	1.42E-006	0.5	0.5	0.5	1.41E-007	0.5	0.25	1	-2.00E-006
0.75	0.75	0	1.72E-006	0.75	0.5	0.5	3.09E-007	0.75	0.25	1	6.18E-007
1	0.75	0	1.17E-006	1	0.5	0.5	7.98E-007	1	0.25	1	2.80E-006
0	1	0	2.12E-006	0	0.75	0.5	-3.84E-007	0	0.5	1	-7.51E-007
0.25	1	0	1.88E-006	0.25	0.75	0.5	3.96E-007	0.25	0.5	1	-1.43E-006
0.5	1	0	2.15E-006	0.5	0.75	0.5	1.31E-006	0.5	0.5	1	-1.50E-006
0.75	1	0	3.04E-006	0.75	0.75	0.5	-1.07E-006	0.75	0.5	1	1.14E-006
1	1	0	1.35E-006	1	0.75	0.5	4.45E-006	1	0.5	1	1.60E-006
0	0	0.25	1.35E-006	0	1	0.5	1.72E-007	0	0.75	1	-7.58E-007
0.25	0	0.25	1.02E-006	0.25	1	0.5	2.69E-006	0.25	0.75	1	1.81E-006
0.5	0	0.25	-2.64E-006	0.5	1	0.5	1.69E-006	0.5	0.75	1	2.09E-006
0.75	0	0.25	-2.18E-006	0.75	1	0.5	1.22E-006	0.75	0.75	1	2.19E-006
1	0	0.25	2.16E-006	1	1	0.5	5.81E-007	1	0.75	1	2.90E-006
0	0.25	0.25	-5.65E-006	0	0	0.75	2.46E-006	0	1	1	2.38E-006
0.25	0.25	0.25	-3.61E-006	0.25	0	0.75	1.42E-006	0.25	1	1	3.30E-006
0.5	0.25	0.25	-1.31E-006	0.5	0	0.75	-1.79E-006	0.5	1	1	1.50E-006
0.75	0.25	0.25	4.50E-007	0.75	0	0.75	-2.34E-006	0.75	1	1	3.32E-006
1	0.25	0.25	2.38E-006	1	0	0.75	2.96E-006	1	1	1	2.44E-006
0	0.5	0.25	-2.29E-006	0	0.25	0.75	-3.69E-006				
0.25	0.5	0.25	-1.19E-006	0.25	0.25	0.75	-3.08E-006				
0.5	0.5	0.25	2.62E-007	0.5	0.25	0.75	-3.00E-006				
0.75	0.5	0.25	1.60E-006	0.75	0.25	0.75	8.47E-007				
1	0.5	0.25	1.00E-006	1	0.25	0.75	2.81E-006				
0	0.75	0.25	2.83E-007	0	0.5	0.75	-7.58E-007				
0.25	0.75	0.25	7.05E-007	0.25	0.5	0.75	-8.35E-007				
0.5	0.75	0.25	2.16E-006	0.5	0.5	0.75	1.79E-007				
0.75	0.75	0.25	9.52E-007	0.75	0.5	0.75	-4.98E-007				
1	0.75	0.25	5.76E-006	1	0.5	0.75	1.58E-006				

Table E.4: Trend on a weekly basis of the average topic weights variance for the FreqLogit model.

c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$	c_π	c_μ	c_ν	$\hat{b}$
0	0	0	4.61E-006	0	1	0.25	1.99E-006	0	0.75	0.75	2.56E-007
0.25	0	0	8.52E-007	0.25	1	0.25	1.42E-006	0.25	0.75	0.75	1.71E-006
0.5	0	0	-2.68E-006	0.5	1	0.25	4.86E-006	0.5	0.75	0.75	2.23E-006
0.75	0	0	-1.74E-006	0.75	1	0.25	3.42E-006	0.75	0.75	0.75	2.86E-006
1	0	0	3.23E-006	1	1	0.25	4.28E-007	1	0.75	0.75	2.96E-006
0	0.25	0	-4.30E-006	0	0	0.5	3.88E-006	0	1	0.75	1.96E-006
0.25	0.25	0	-2.92E-006	0.25	0	0.5	1.19E-006	0.25	1	0.75	1.24E-006
0.5	0.25	0	-2.59E-006	0.5	0	0.5	-2.43E-006	0.5	1	0.75	3.81E-006
0.75	0.25	0	-2.12E-007	0.75	0	0.5	-2.59E-006	0.75	1	0.75	2.56E-006
1	0.25	0	2.08E-006	1	0	0.5	2.60E-006	1	1	0.75	4.16E-006
0	0.5	0	-3.78E-006	0	0.25	0.5	-4.95E-006	0	0	1	2.21E-006
0.25	0.5	0	-2.02E-006	0.25	0.25	0.5	-3.58E-006	0.25	0	1	1.37E-006
0.5	0.5	0	6.92E-007	0.5	0.25	0.5	-1.22E-006	0.5	0	1	-2.42E-006
0.75	0.5	0	-4.22E-007	0.75	0.25	0.5	9.41E-007	0.75	0	1	-2.73E-006
1	0.5	0	4.08E-006	1	0.25	0.5	2.31E-006	1	0	1	3.64E-006
0	0.75	0	6.02E-007	0	0.5	0.5	-3.03E-006	0	0.25	1	-3.95E-006
0.25	0.75	0	-7.13E-007	0.25	0.5	0.5	-2.85E-007	0.25	0.25	1	-3.57E-006
0.5	0.75	0	8.40E-007	0.5	0.5	0.5	6.43E-007	0.5	0.25	1	-3.12E-006
0.75	0.75	0	1.93E-006	0.75	0.5	0.5	4.12E-006	0.75	0.25	1	1.20E-007
1	0.75	0	1.66E-006	1	0.5	0.5	2.90E-006	1	0.25	1	1.35E-006
0	1	0	4.08E-006	0	0.75	0.5	6.48E-007	0	0.5	1	-2.92E-006
0.25	1	0	1.27E-006	0.25	0.75	0.5	1.40E-006	0.25	0.5	1	-1.25E-006
0.5	1	0	3.23E-006	0.5	0.75	0.5	2.51E-006	0.5	0.5	1	9.80E-007
0.75	1	0	5.28E-006	0.75	0.75	0.5	-2.50E-007	0.75	0.5	1	-1.18E-006
1	1	0	4.72E-006	1	0.75	0.5	1.78E-006	1	0.5	1	3.64E-006
0	0	0.25	5.01E-007	0	1	0.5	2.80E-006	0	0.75	1	4.52E-007
0.25	0	0.25	8.95E-007	0.25	1	0.5	4.23E-006	0.25	0.75	1	1.77E-006
0.5	0	0.25	-2.03E-006	0.5	1	0.5	2.99E-006	0.5	0.75	1	1.76E-006
0.75	0	0.25	-4.62E-006	0.75	1	0.5	3.05E-006	0.75	0.75	1	3.81E-006
1	0	0.25	3.51E-006	1	1	0.5	2.77E-006	1	0.75	1	4.05E-006
0	0.25	0.25	-4.42E-006	0	0	0.75	2.76E-006	0	1	1	3.81E-006
0.25	0.25	0.25	-3.36E-006	0.25	0	0.75	7.50E-007	0.25	1	1	2.87E-006
0.5	0.25	0.25	-2.45E-006	0.5	0	0.75	-3.12E-006	0.5	1	1	3.03E-006
0.75	0.25	0.25	1.15E-006	0.75	0	0.75	-2.22E-006	0.75	1	1	4.96E-006
1	0.25	0.25	2.73E-006	1	0	0.75	4.33E-006	1	1	1	3.54E-006
0	0.5	0.25	-1.87E-006	0	0.25	0.75	-3.45E-006				
0.25	0.5	0.25	-4.04E-007	0.25	0.25	0.75	-4.33E-006				
0.5	0.5	0.25	-7.11E-007	0.5	0.25	0.75	-1.80E-006				
0.75	0.5	0.25	1.30E-006	0.75	0.25	0.75	2.64E-007				
1	0.5	0.25	2.88E-006	1	0.25	0.75	2.94E-006				
0	0.75	0.25	-1.44E-006	0	0.5	0.75	-1.38E-006				
0.25	0.75	0.25	6.35E-007	0.25	0.5	0.75	-1.23E-006				
0.5	0.75	0.25	1.86E-006	0.5	0.5	0.75	-9.02E-007				
0.75	0.75	0.25	1.17E-006	0.75	0.5	0.75	1.88E-006				
1	0.75	0.25	2.00E-006	1	0.5	0.75	1.04E-006				

Table E.5: Trend on a weekly basis of the average topic weights variance for the FreqRandom model.

Appendix F

Bibliography

[1] Iab internet advertising revenue report 2011. Technical report, IAB, 2012.

[2] Internet en chile 2011. Technical report, IAB Chile, 2012.

[3] http://www.facebook.com/zuck/posts/10100518568346671, Retrieved October 12th, 2012.

[4] http://www.iab.cl/2012/08/14/falabella-es-la-primera-empresa-chilena-en-alcanzar-1-millon-de-seguidores-en-facebook/, Retrieved October 12th, 2012.

[5] http://www.nytimes.com/2012/02/19/books/review/how-an-egyptian-revolution-began-on-facebook.html, Retrieved October 12th, 2012.

[6] http://www.plexilandia.cl/faq.html, Retrieved October 12th, 2012.

[7] http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919, Retrieved October 14th, 2012.

[8] http://jgibblda.sourceforge.net/, Retrieved October 27th, 2012.

[9] http://www.cs.waikato.ac.nz/ml/weka/, Retrieved October 27th, 2012.

[10] http://www.mysql.com/downloads/connector/j/, Retrieved October 27th, 2012.

[11] http://www.plexilandia.cl/foro, Retrieved October 28th, 2012.

[12] Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *International Review of Economics*, 1(1):pp. 3–49, 2011.

[13] J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integreated theory of the mind. *Psychological Review*, 111(4):pp. 1036–1060, 2004.

[14] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference*,

January 2012.

[15] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[16] Frank M. Bass. A new product growth model for consumer durables. *Management Science*, 16(5):pp. 215–227, 1969.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:pp. 993–1022, 2003.

[18] Rafal Bogacz, Marius Usher, Jiaxiang Zhang, and James L McClelland. Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B*, 362:1655–1670, 2007.

[19] Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology*, 2:article 22, 2011.

[20] James Coleman, Elihu Katz, and Herbert Menzel. The diffusion of an innovation among physicians. *Sociometry*, 20(4):253–270, December 1957.

[21] Chris Davis and Tjark Freundt. What marketers say about working online. In *mckinseyquarterly.com*, November 2011.

[22] Emmanuel Dion. *Invitation à la théorie de l'information*. Éditions du Seuil, 1997.

[23] Roxane Divol, David Edelman, and Hugo Sarrazin. Demystifying social media. *McKinsey Quarterly*, April, 2012.

[24] David Easley and Jon Kleinberg. *Networks, Crowds and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[25] P. Erdös and A Rényi. On the evolution of random graphs. In *Publication of The Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[26] Peter M. Guadagni and John D. C. Little. A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2(3):pp. 203–238, 1983.

[27] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW 2005*, 2005.

[28] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010.

[29] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physicial Character*, 115:pp 700–721, 1927.

[30] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of

epidemics. ii. the problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physicial Character*, 138:pp. 55–83, 1932.

[31] W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. iii. further studies of the problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physicial Character*, 141:pp. 94–122, 1933.

[32] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):pp. 1–39, 2007.

[33] Peter L.T. Pirolli. Power of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42:pp. 33–40, 2009.

[34] Peter L.T. Pirolli and Wai-Tat Fu. Snif-act: a model of information foraging on the world wide web. In *9th International Conference on User Modeling*, 2003.

[35] Sidney Redner. Py896 course handbook: Fundamental kinetic processes. Chapter 6 can be downloaded at http://physics.bu.edu/ redner/896/spin.pdf.

[36] Everett M. Rogers. *Diffusion of Innovations*. The Free Press, 1962.

[37] Everett M. Rogers and F. Floyd Shoemaker. *Communication of innovations: a cross-cultural approach*. The Free Press, 1971.

[38] Pablo Roman. *Web User Behavior Analysis*. PhD thesis, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, 2011.

[39] Bryce Ryan and Neal C. Gross. The diffusion of hybrid seed corn in two iowa communities. *Rural Sociology*, 8:pp. 15–24, 1943.

[40] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[41] Gabriel Tarde. *The Laws of Imitation*. Henry Holt and Company, 1903.

[42] Gabriel Tarde. *Les lois de l'imitation*. Librairie Félix Alcan, 7 edition, 1921.

[43] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, December 1969.

[44] Marius Usher and James L McClelland. The time course of perceptual choice: the leaky, competing acumulator model. *Psychological Review*, 108(3):550–592, 2001.

[45] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.