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

MODELO NEUROFISIOLÓGICO PARA LA DIFUSIÓN DE INFORMACIÓN EN REDES SOCIALES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

PABLO ANDRE CLEVELAND ORTEGA

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
PABLO ROMÁN ASENJO
FELIPE AGUILERA VALENZUELA
DENIS SAURÉ VALENZUELA

SANTIAGO DE CHILE
2018

RESUMEN

El creciente uso de los servicios de internet, particularmente de las redes sociales (OSN) ha generado una gran oportunidad para entender mejor el comportamiento de los usuarios como también de los flujos de información. A pesar de que la modelación de los flujos de información no es un tema nuevo, sí es de mucha dificultad y gracias a la aparición de OSNs y comunidades virtuales de practica (VCoPs) es que ha resurgido como tema, gracias a la disponibilidad de data. Sin embargo, la mayoría si no todos los estudios revisado solo modelan a un nivel macroscópico, donde los grandes números absorben comportamientos indeseados y así se reportan buenos resultados. Nuestra hipótesis es que es posible modelar la difusión de información a nivel microscópico mediante un modelo derivado de la neurofisiología.

El objetivo principal de este trabajo es desarrollar e implementar una metodología para predecir el intercambio de información entre usuarios a un nivel microscópico usando el contenido de texto mediante técnicas de Text Mining, con el fin de apoyar el proceso de administración de una VCoP.

Para ello se propone una metodología que combina dos procesos Knowledge Discovery in Databases (KDD) y SNA y fue aplicada sobre una VCoP real llamada Plexilandia. En la etapa de KDD se efectuó la selección, limpieza y transformación de los posts de los usuarios, para luego aplicar una estrategia de reducción de contenido Latent Dirichlet Allocation (LDA), que permite describir cada post en términos de tópicos. En la etapa de SNA se aplicó un modelo neurofisiológico de toma de decisiones adaptado a preferencias de texto para predecir la formación de arcos entre hilos y usuarios usando la información obtenida en la etapa anterior.

Los resultados de los experimentos muestran que es posible predecir con un alto porcentaje de éxito, 65 a 80% cuando hay poco ruido y 40 a 60% cuando existe elevado ruido, las interacciones entre usuarios basándose en la similaridad de los textos producidos por ellos. Esto permite vislumbrar la forma en que se difundirá un mensaje e identificar a usuarios que potencialmente estén interesados en un hilo.

ABSTRACT

The growing use of Internet services, particularly social networks (OSN) has generated a great opportunity to better understand the behavior of users as well as information flows. Although the modeling of information flows is not a new issue, it is very difficult and thanks to the emergence of OSNs and virtual communities of practice (VCoPs) is that it has resurfaced as a theme, thanks to the availability of data. However, most if not all the studies reviewed only model at a macroscopic level, where large numbers absorb unwanted behavior and thus good results are reported. Our hypothesis is that it is possible to model the dissemination of information at the microscopic level through a model derived from neurophysiology.

The main objective of this work is to develop and implement a methodology to predict the exchange of information between users at a microscopic level using the text content extracted by means of Text Mining techniques, to support the administration process of a VCoP.

To this end, a methodology is proposed that combines two Knowledge Discovery in Databases (KDD) and SNA processes and was applied to a real VCoP called Plexilandia. In the KDD stage, the selection, cleaning and transformation of user posts was carried out, followed by a content reduction strategy using Latent Dirichlet Allocation (LDA), which allows to describe each post in terms of topics. In the SNA stage, a neurophysiological decision-making model adapted to text preferences was applied to predict the formation of arcs between threads and users using the information obtained in the previous stage.

The results of the experiments show that it is possible to predict with a high percentage of success (65 to 80 % when there is little noise and 40 to 60 % when there is high noise) the interactions between users based on the similarity of the texts produced by them. This allows to glimpse the way in which a message will be spread and to identify users who are potentially interested in a thread.

For Myriam

Acknowledgements

First, I want to thank my fiancé Karla, who has been my motivation and source of strength during this stage of my life.
I also want to thank my advisor Dr. Sebastián Ríos and co-advisor Pablo Román, the main supporters of my master's activities, for their patience, dedication, motivation and encouragement. Thanks for giving me this opportunity.
My family for all their support on this journey.
My godparents for all their help in moving to Santiago and during my stay here.
Felipe Aguilera and the team at "Deta Consultores"for giving me the opportunity to participate in an interesting project and providing me with many tools that helped me improve my skill set.
I want to thank CONICYT for awarding me with the "Magíster Nacional"grant which allowed me dedicate in full to my studies.
Finally, this research would not be possible without to the data base of the site `http://www.plexilandia.cl/foro`, therefore, I would like to thank José Ignacio Santa Cruz and Plexilandia's Community.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction & Motivation

In this chapter the main motivation to study information diffusion on online social networks is reviewed followed by the statement of the main and specific goals of this thesis. After that the methodology used for the development of this research is presented, and finally a brief description of the remaining chapters is given.

1.1 The Information Diffusion Problem

Massive use of Internet services such as social networks allowed people to communicate and interact with each other, without concern for their geographic location. It is possible for someone to find people to converse with, people with common interests, to help others in certain problems, to share information, to participate in discussions, etc. These activities changed the use of the computer from an individual activity to a collective one, this in turn has been responsible for creating different links of interaction and cooperation with other people [16]. All the above has contributed to an increasing relevance of Internet in our daily lives.

The importance of Internet has led to the emergence of new social institutions [9, 16]: Online Social Networks (OSN), Virtual Communities (VC), along with other types of social entities. Although, based on existing ones, they possess specific characteristics [11] that need to be considered while performing a study of them. These differences are due to the use of a different medium than the face-to-face interaction, which generates many real-world social rituals [10] that do not exist or are limited in the virtual world [14].

To support the latter new social structures, it is possible to use different technologies. For example, one may use a wiki system or a forum, a blogging system, a messenger system, e-mail lists, among many other. Also, it is possible to use

a combination of more than one of these technologies to support a community or an online social network. Usage of these technologies not only enriches the information shared it also accelerates the diffusion process.

The broadly spread usage of these technologies has generated an opportunity to study the way users interact with each other, the influence they have on others, the way a certain piece of information spreads across the network, etc. Although, it was known that these problems are important they were very difficult to measure before OSN. We focus on the problem of understanding how information spreads or is shared within an OSN. This problem is of particular interest and finds applications on many areas, such as: generating a viral marketing campaign successfully, political campaigns, being able to detect malicious or inaccurate rumors and even preventing terrorist attacks, measuring and tracking social events like revolutionary waves, disease spreading, among many others.

OSNs can be separated into different groups according to the features they have, for instance Facebook, Sina-weibo and Twitter possess characteristics (friend or follower relationships) that allow to reconstruct the social network graph in a straight forward fashion. It can also be said that social relationships play a major role on these OSNs. Furthermore, the information shared in these networks is often lighter in content and profundity and very influenced by social interactions. This leads us to select another type of OSN, Web Forums. These have many desirable characteristics for studying information diffusion, for example, due to the lack of friendship or follower relationships the main focus of users when surfing a web forum is the content of the conversations (threads) contained in it, thus making a forum content-driven. Another desirable characteristic is they provide a truly open and freely accessible platform for information diffusion as noted in [69] because anyone can begin a new thread or participate in an existing one facilitating opinions to be freely formed and shared. Besides, information posted in OSN like Facebook has the problem that it is filtered by a recommendation algorithm first, and it is well known that this kind of algorithms, based on browsing history, restrict the content presented and browsed by the user, which of course, will provide biased information to our study. We consider information content to be of utmost importance to understand the diffusion process which is the reason why this research will focus on web forums.

Studies performed in OSNs can be separated on whether they describe macroscopic, mesoscopic or microscopic behavior of the structure. Examples of the first two are studies of density of the network representation of the community or other global network characteristics like power law degree distribution that can be associated to the empirical observation that most of the content is produced by a

small percentage of web users, like mentioned in the work of Baeza-Yates [74]. As we try to delve deeper into the mechanisms of diffusion, trying to understand the role of each agent, the problem gets increasingly complex mainly due to the exponentially increasing amount of possible interactions. In this work our interest is to model agent decision making in the diffusion process, so we fall in the last category. We can clearly see the increase in complexity if we take for instance a typical problem that falls into the microscopic level category and one that falls into the mesoscopic-macroscopic. For instance, take the problem of obtaining the set of arcs formed in a network of N nodes for the micro level and the problem of obtaining the edge density for the macro-meso level. If there are k arcs in the network the probability of getting the right set is:

$$P(\text{Guessing all k arcs}) = \frac{1}{N+1}\frac{1}{\binom{N}{k}} \tag{1.1}$$

And the probability of getting the right edge density is:

$$P(\text{Guessing edge density}) = \frac{1}{N+1} \tag{1.2}$$

Where $P(\text{Guessing all k arcs}) \ll P(\text{Guessing edge density})$ for most cases. It is also worthy of note that with the results of a micro model you can later obtain results at a meso or macro level, however it is not possible the other way around.

Recapitulating, our work will focus on modeling agent decision making in the diffusion process on web forums, with a content-driven approach and a web administrator's standpoint. We believe that knowing the conversations which might be of interest to a particular user of great relevance to a web forum administrator. For instance, it allows him to recommend newly created discussions in the forum. Also knowing the details of information diffusion within the forum permits him to have better judgment while performing his administrator duties.

The main problematics we faced while trying to model information diffusion are the following

- The social networks are often extremely large and complex, furthermore if you add the information (text content) component to the equation the problem becomes even harder to model and process.
- Most techniques focus only on a small portion of the sources of information available. For instance, on the one hand, most Social Network Analysis

(SNA) algorithms and statistics perform an automated analysis to gather valuable information about community structure based on relationships between community members. On the other hand, the data mining approach, in particular web mining (WM), which is the application of data mining algorithms to web generated data, where the structure of social interactions is lost but allows us to find interesting patterns of texts in members posts or navigation patterns [31, 34, 38, 43].

- The vast majority of models focus on obtaining results at an aggregated level, e.g. amount of people with interest in a certain topic. However, either it's not possible to adapt them to the microscopic level or when it is possible the results are no longer at an acceptable quality level.
- There is a lack of standardization in the field, i.e., most works about diffusion are made in a way such that it is not possible to make comparisons. In the work made by Guille et al [53] an attempt was made to standardize the works with Sondy an on-line platform that allows researchers to perform most SNA algorithms. However, it was not adopted by the community of researchers.
- There is no documentation about the many steps that need to be taken while on the process of developing a model for information diffusion which makes the process itself a lot more prone to error.

Our problem is complex from the point of view of modeling and data processing. But at the same time is extremely relevant to be able to understand the dynamics of human behavior in online social networks, which today have already established themselves as a fundamental mechanism in many areas of daily life.

1.2 Research Hypothesis

The fundamental intuition underlying this work is that it is possible to apply a neurophysiological model of decision making to model information diffusion at a microscopic level and obtain quality results. Specifically, one research hypothesis is formulated:

R. H. (**capability of Leaky Competing Accumulator (LCA) of modeling information diffusion at a microscopic level**) it is possible to use a customized LCA model, as an underlying mechanism of decision, to model agent interactions with regards to information content on a web forum

1.3 Goals & Results

1.3.1 Main Objective

The main objective of this work is to develop and implement a methodology to predict the exchange of information between users at a microscopic level using the text content extracted by means of Text Mining techniques, to support the administration process of a Web Forum.

1.3.2 Specific Objectives

1. Review the literature of information diffusion to paint an image of the current state and create a benchmark of models
2. Develop a methodology for information diffusion modeling in web forums using the LCA model.
3. Implement a model that allows us to capture the decision-making process of the agents participating in the network

1.3.3 Expected Results

Some of the expected results of this thesis are:

1. Chapter 1 and 2 of this thesis.
2. A methodology implemented, tested and evaluated.
3. A model that in average obtains 50% or more in F-measure score
4. A framework of software that can be used by other researchers in the area.

1.4 Methodology

The work methodology is based on the process of Knowledge Discovery in Data Bases (KDD) and SNA. We will work with the data obtained from a real Web Forum. For the development of this thesis previous work data of search of key members in a Web Forum made by Álvarez [34] will be used. Subsequently, using the aforementioned data and text mining strategy, the inputs of the LCA model will be computed. Next, LCA model calibration will be performed through the

use of the genetic algorithm and then simulations of the network will be run as shown in Fig. 1.1. The methodology used for the development of this thesis is

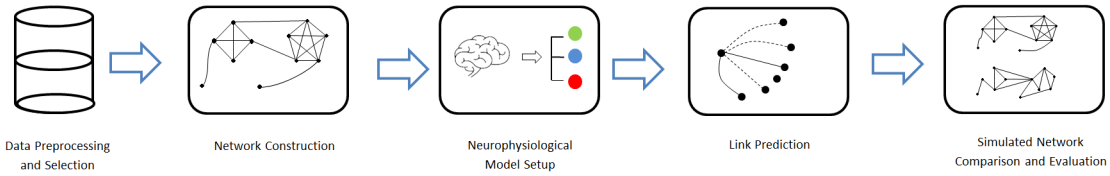


Figure 1.1: Proposed Methodology for Modeling Information Diffusion in Web-Forums.

carried out in the following steps:

1. **Previous advances in the area and Information Diffusion**
   For the development of this thesis work, a brief introduction of SNA applications will be made, describing their study areas. Subsequently a review of the different methods that address the problem of information diffusion based on SNA and some of the latest advances in the area will be made. In addition, some methods that seek to combine text mining that use topic models or concepts to enhance link prediction in networks, will be reviewed.

2. **Graph Representations**
   Representations of social networks may be different depending on the characteristics of the SNA problem to be addressed. That is why it is necessary to use strategies that allow the construction of graphs that capture all the information contained in a social network and allow the subsequent analysis with SNA tools with those that address the problem of information diffusion.

3. **Algorithm customization**
   LCA model needs to be integrated and customized to the link prediction problem using text content as the information available of the network. Furthermore, an optimization scheme must be implemented in order to calibrate the model parameters.

4. **Network Simulation**
   The algorithm will be tested over a real Web Forum. First, using the data

from previous work done by Ríos et al. [31, 34, 41, 48, 71] modification will be made to extract the network graphs and node features needed for the for the subsequent implementation of the algorithm. Next, the customized algorithm will be run and the simulated networks of the forums will be obtained in a fashion such that it allows the evaluation of the algorithm and, thus, of the methodology.

5. **Results analysis and conclusions**
   Once the proposed methodology is implemented, its results will be evaluated by the use of 4 different metrics. Finally, the conclusions of each of the stages described above will be presented.

1.5 Thesis Structure

In the following chapter a review of the most relevant bibliography regarding information diffusion is discussed.

In chapter 3 the methodology proposed to implement the information diffusion model including

Afterwards in chapter 4 a real Web Forum Plexilandia is described as well as the details of the conducted experiment and its results.

Finally, in chapter 5 the main conclusions of this thesis are presented, including the main contributions of this work as well as the next investigation lines and future work.

Chapter 2

Related Work

There is a huge amount of research that deals with social networks. Some of this work is centered in discovering communities within the network [41, 43, 58], influencers and key member discovery [18, 26, 34, 36, 40, 60, 61, 71], macro-structural properties of networks [12] Other studies have focus on describing the evolution of certain networks like [30, 31, 54]. for example, Jianwei Niu et al [54] a descriptive analysis of a Facebook-like Chinese OSN named Renren is conducted where macroscopical structural and evolutionary properties of the network are described and seem to match the results of previous work conducted on similar networks.

We will focus on works that model information diffusion and link prediction as we will take a combined approach in our proposed model. A more detailed explanation of them is given below.

2.1 Information Diffusion

We start this review by defining diffusion which refers to the process whereby a phenomenon of interest (e.g., information, innovation, or disease) spreads from one to another as stated in [69]. Guille et al [52] developed a survey of previous research regarding information diffusion in which they present some basic definitions and classify previous work according to their respective contribution and novelty. They also describe the different approaches used to model information diffusion and define three questions which are at the core of the field, namely:

1. Which pieces of information or topics are popular and diffuse the most?

2. How, why and through which paths information is diffusing, and will be diffused in the future?

3. Which members of the network play important roles in the spreading process?

In this research we try to answer the second question. As noted in [52], information diffusion has been approached in many ways. Therefore, we will make a classification of proposed models into families as shown in Fig. 2.1 in order to get a better grasp of each of these approaches and to be able to present previous research in an orderly fashion.

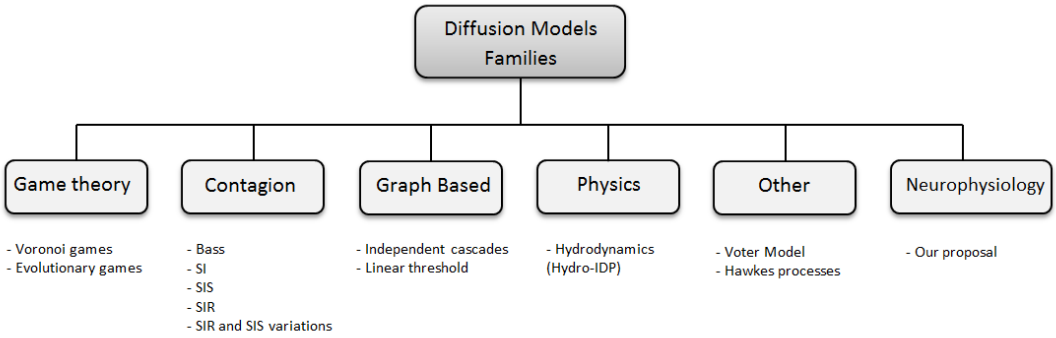


Figure 2.1: Diffusion Models Classification

2.1.1 Game Theory Models

One approach frequently used to deal with information diffusion is based in game theory, in particular, Voronoi games in graphs [21, 32, 49, 55, 56]. For instance, in Nora Alon et al [32] a competitive game-theoretic model for diffusion is shown which may be useful for understanding situations where competing products are advertised via viral marketing campaigns. They also claim a relation between the diameter of a given network and the existence of a pure Nash equilibria in the game which was later corrected by Reiko Takehara et al [49]. After that Lucy Small and Oliver Mason [56] showed that the claim holds true if the underlying graph is a tree and later in [32] they extended their results by extending the model to a graph that follows the iterated local transitivity model for social networks. They showed that for 2 competing agents, an independent Nash Equilibrium on the initial graph remains a Nash Equilibrium for all subsequent times. In [57] they use

the evolutionary game over uniform degree networks to model users' information forwarding strategies, i.e. to forward the information or not. They test this model over Twitter hashtags dataset validating their proposed model.

2.1.2 Contagion Models

Another family of models frequently used when dealing with information diffusion are epidemic spreading or contagion models. The first model we will refer to is the Bass Model, in [3] he presents a model for the sales of a new product as a function of time that originates from contagion models. The most common models found in this category are: susceptible-infected (SI), susceptible-infected-susceptible model (SIS) and susceptible-infected-recovered model (SIR), among other variations.
Ye Sun et al in [59] studied the impact of using weighted edges to represent multi-role relations on a network when studying two representative macroscopic metrics, outbreak threshold and epidemic prevalence. They conducted experiments using 2 sets of weight distributions, Uniform and Poisson distribution, on a small-world network and a scale free network. They tested both SIS and SIR models. Their results show good agreeance with theorical results except that the simulation results show that the weight distribution effect is very weak. The main result of this work is that, on fully mixed networks, the weight distribution on edges would not affect the epidemic results once the average weight of the whole network is fixed. Saxena et al [62] proposed a model with hierarchical probabilities of infection across edges that depend on the relative position of the end-nodes on the network (core or periphery). They tested the model with twitter data obtaining similar behavior patterns of diffusion.

Kubo et al. [23] applied the SIR model to capture human behavior in a virtual community, especifically "2 channel"the biggest Japanese open anonymous Bulletin Board System (BBS). Jiyoung Woo et al. in [44] implement the SIR model at a topic level to model violent topic spreading. They test this model on Ummah dataset from the dark web forum portal developed by the artificial intelligence lab of the university of Arizona. Later, in [50] they test and modify the previous SIR Model to incorporate the effect of online news, proposing the event-driven SIR model. This model captures the effect of online news on the infection rate, population growth and infected group growth. They tested this model on Yahoo! Finance-Walmart message board and use Walmart-related news from the Wall Street Journal. Next, in [69] they propose a contagion (epidemic) model SIR to model information diffusion in Web-Forums due to similar patterns

in the spread of information and social contagion processes. They build up on previous work [23, 44, 50] changing the view from post-level to author-level information diffusion. In Xiong et al. [51] they propose another variation of a contagion model, the susceptible-contacted-infected-refractory (SCIR) model. They tested the model by numerical simulations. In [68] Qiu et al. incorporate a forgetting and refuting mechanism into the SIR model to describe rumor spreading more accurately. They tested their model in both numerical simulations and Renren OSN.

Other approach is presented in [47] where they present a model that incorporates the effect of external sources of influence into the infection process, which they model with hazard functions, complementing the information diffusion description. They test this model on synthetic data and on Twitter. In their experiments they concluded that about 70% of diffusion on Twitter can be attributed to be caused by a network effect and the rest (30%) is due to external sources such as online news, Facebook, etc.

2.1.3 Graph Models

Mainly 2 models fall into this category, namely Independent Cascades (IC) [13] and Linear Threshold (LT) [4]. As described in [52] they both assume the existence of a static graph structure underlying the diffusion and focus on the structure of the process. IC associates a probability to each edge that represent the chance of information being diffused. LT defines a threshold for each user (node) and an influence degree for each edge. Information diffuses, or a node becomes active in LT if the sum of influence of active neighbors of a user surpasses his threshold. We will not delve deeper into these models because our goal is to model diffusion on web forums where there is no explicit graph.

2.1.4 Others

In this category we classify all proposed models that do not fit in the aforementioned categories. For example, the voter model. In [21] they use this model to represent the diffusion of opinions in a social network. However, their aim is to solve the spread maximization set problem which differs from our goal. Another model that falls into this category is one that uses multidimensional Hawkes processes which are a class of self or mutually exciting point process models. They capture underlying user influence with this model and validate the model by testing it on synthetic and Twitter data sets. Hu et al. tried a different approach [70]

in which they implement a non-parametric hydrodynamics model adapted to information diffusion (Hydro-IDP) by correlating the characteristics of the fluid-density flow evolution in the physical space-time with that of information diffusion in the cyber space-time. For this purpose, they come up with the analogy between initial source energy, initial source radius, initial flow velocity and information popularity, publisher's influence, diffusivity of the social platform respectively simultaneously defining new features of interest to be studied in a social network. They test their model on data of China's OSN Sina-weibo. Lee et al. [42] use spatial interaction models typically used in the field of economics and economic geography to study the relationship between distance and acquaintanceship between university students using data from StudiVZ.

2.1.5 Previous Work Classification

As we can notice in Table 2.1 most of the reviewed previous research focuses on OSNs like Facebook and Twitter both of which possess the particularity that an explicit social network graph can be easily extracted by the use of friendship or follower relationships respectively. In turn this makes the models proposed on these OSNs heavily rely upon this information and it is not clear if these proposed models can be applied to OSNs in which this information is lacking. As mentioned before, our goal is to study web forums so we put emphasis on the works that are applied to them such as [23, 44, 50, 69]

2.2 Link Prediction

The link prediction problem consists in being able to predict relationships in a network. The clear majority of research done in regard to this problem uses local network structure in order to obtain the likelihood of two nodes of the network forming a link. The problem accepts two classic definitions: one is related to the evolution of the network in which the question we seek to answer is whether the current state and topology of the network can be used to predict the future state and topology, i.e., can future links be predicted. The other definition refers to a situation where one is missing information about the network, namely, some of the links, and the question we must answer is if it is possible to infer the missing links using the information we have available. In this research we work with the second definition. An extensive survey was done in [39] where they examined various approaches to the problem like Feature models, Bayesian graphical models and the linear algebraic approach, comparing model complexity, prediction performance,

Table 2.1: Previous Work Classification

Reference	**OSN**				**Model**			**Level**		
	Facebook or Twitter-like	Simulated network	Web-Forum-like	Theoric (no data)	Deterministic	Probabilistic	None	Macroscopic	Mesoscopic	Microscopic
C. Jiang et al (2014) [57]	✗				✗			✗		
N. Alon et al (2010) [32]				✗	✗					✗
Y. Sun et al (2014) [59]		✗				✗		✗		
A. Saxena et al (2015) [62]	✗					✗			✗	
S. Myers et al (2012) [47]	✗					✗			✗	
Y. Hu et al (2017) [70]	✗				✗				✗	
L. Small and O. Mason (2013) [55]				✗	✗					✗
F. Xiong et al (2012) [51]		✗			✗			✗		
J. Woo and H. Chen (2012) [50]			✗		✗				✗	
J. Woo and H. Chen (2016) [69]			✗		✗				✗	
J. Woo et al (2011) [44]			✗		✗				✗	
X. Qiu et al (2016) [68]	✗				✗			✗		
M. Kubo et al (2007) [23]			✗		✗			✗		
Proposed Model			✗			✗				✗

among others. In [73] the link prediction problem is approached as an optimization problem with cardinality constraints and ITERCLIPS frameworks is proposed to obtain solutions. In [65] they tackle the problem of evaluation to standardize the metrics used and make results comparable between researches. The works that we could find that resemble our approach the most are the following: In [72] they propose a new model that uses a set of available meta-paths to estimate the link likelihood. They test this model on a Bibliography network where metadata can be obtained. The results obtained by this model seem very promising. In [67] they use text content diffusion, as in being exposed to a post originally submitted by a unrelated user, in order to improve link prediction. Finally, in [24] a modification

of LDA is proposed where they include text content, mainly research keywords, to improve prediction of future collaborations between authors. We have not found a work that resembles more our approach to the problem.

Chapter 3

Proposed Methodology

As mentioned before the main question of this work is how to model agent decision making in the diffusion process on web forums, with a content-driven approach and a web administrator's standpoint. However, to accomplish that, we need to address certain issues first as we can observe in Fig. 3.1.

3.1 Data Selection and Preprocessing

The first problem we must face is preprocessing the data of the forum in order to be able to use it for the model. There are several ways to obtain data from social networks. In general, it will depend on the problem we are approaching to determine the appropriate method for this process. Culotta [20] presents a methodology for obtaining data from a social network from e-mails and Ríos et al. [64] presents the basic steps for extracting data from a VCoP, which usually bases its operation, on forums systems such as VBuletin[1], MyBB[2] and phpBB[3], among others. A forum in the context of the web, refers to a virtual place in the space where members interact, discuss ideas, share and generate knowledge. In general, the topics within a forum are arranged in a hierarchical way, with different categories according to the interest of the members that frequent it. In the case of VCoP, the categories are directly related to the purpose of the community. Each conversation in the forum, within the categories, is called Thread or discussion topic and in them the members present their opinions and discuss around a central idea. Each message between members made within a Thread is called post, which

[1]https://www.vbulletin.com [Access date April 10, 2018]
[2]http://www.mybb.com [Access date April 10, 2018]
[3]http://www.phpbb.com [Access date April 10, 2018]

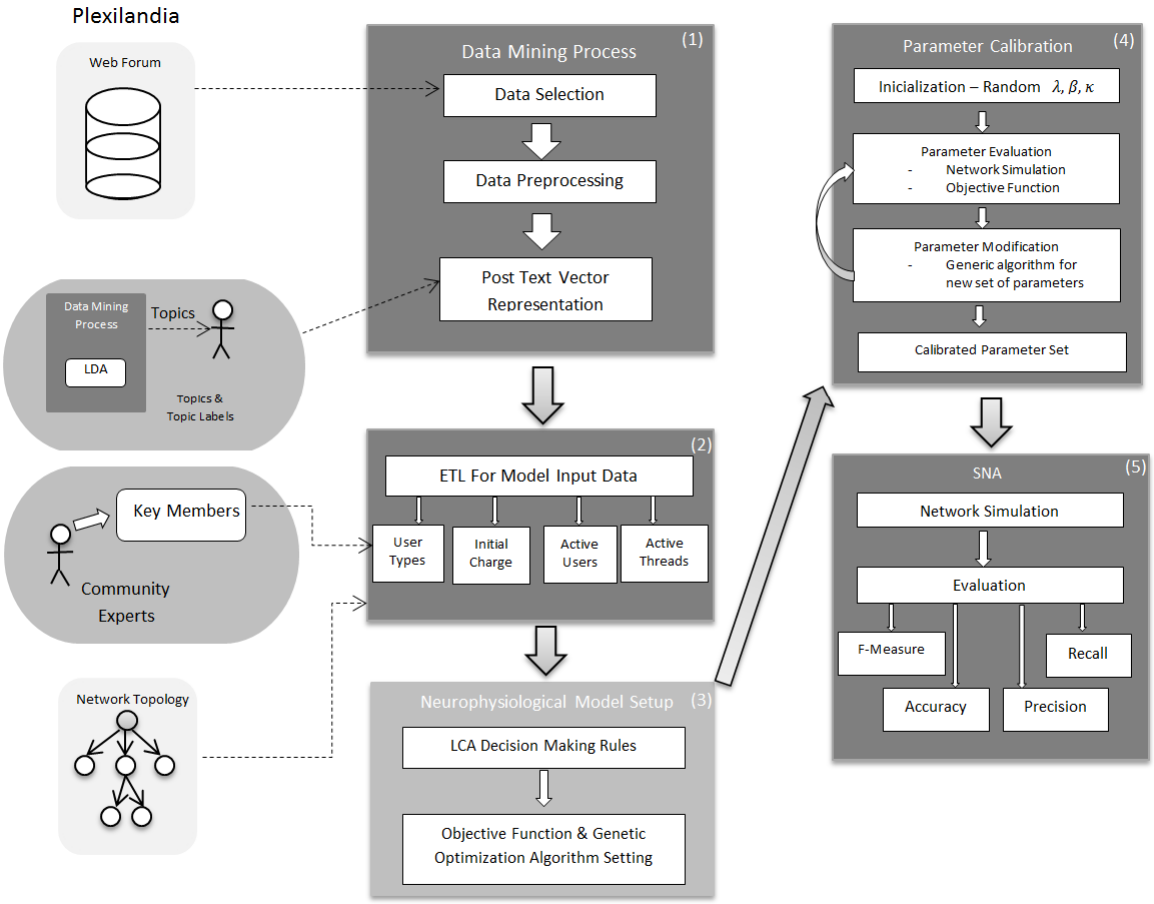


Figure 3.1: Framework for the LCA model.

is the basic unit within a forum. A discussion topic begins with a post from a user in the community, which in general, contains a question or the presentation of an idea that you want to discuss. Then, the different members of a VCoP, related to the user or the topic of discussion, make their posts to be debated and thus generating knowledge about the central theme of the community.

Each post is composed of some basic elements such as the user identifier (ID), which allows to know which members of the community are interacting in the discussion; the content of the post, which depending on the forum can be text, images, links to other pages, videos, etc.; and the information of the forum system, such as the date of creation of the post, the Thread and the category to which

it belongs, among others. According to the methodology exposed by Ríos et al. [64], the basic elements just mentioned, and the content will be selected as data only in text format available in the forum. For the cleaning of the data, Ríos et al. [64] indicates that the first filter should be made regarding the replies (quotes) of other posts. In a forum, a user can respond to another post by creating a new message with the copy of the cited post plus the additional text exposed by this new user. Therefore, it is necessary to delete the replicated part of the new post and only store the new text input.

Then you must transform the acronyms or abbreviations, eliminate spelling errors and all the elements of the posts that make these not comparable. This process is carried out by two methods: The first is a process called stemming which involves the transformation of each word to its root, the modification of the plural words into singular and the change of all the phrases or words that do not represent a contribution to the data of the text, assigning a representative term as "unusable" or "miscellaneous". The second method called stop words filter where all the words that do not contribute information to the network are deleted, such as articles, pronouns and "unusable" words. The objective of the aforementioned methods is, on the one hand, to make the posts comparable to evaluate if both speak of the same and on the other hand, to reduce the number of words used to make the comparison. However, the number of resulting words could be too large to relate the posts only based on useful words. To reduce the content of the vocabulary used in the network and to be able to adequately compare each post, the methods of text processing will be used.

3.1.1 Text processing based on text mining

After the selection, cleaning and transformation of the data, it is necessary to reduce the large number of existing words in the forum to make a successful comparison of the content of the network. However, you cannot perform any content reduction strategy, you must use methods that allow you to extract the central ideas of each post. One way to do this is to define a series of concepts, topics, or categories and classify each message according to its proximity to each of the categories. When carrying out this process properly, the posts could be described by a reduced number of concepts, which would be very useful to make comparisons between them. The problem is that, the different categories to be formed will depend on the structure of the forum, the objective that it pursues, and the topics dealt within it. To solve the above, the use of topic or concept models is proposed as a strategy for reducing the content of the network. The models of

topics and concepts allow to describe the thematic content of documents without prior classification [25], however, it is necessary to represent the data obtained in the previous stage in a way that allows its application. Next, the notation and representation of the data will be presented to implement the topic models, followed by a strategy for the reduction of content and classification of posts.

3.1.2 Representation of the data

Each of the messages obtained in section 3.1 contains a series of words that characterize the post. If all the different words of all messages are taken, then the complete vocabulary used by the network will be obtained.

Let $\mathcal{V}$ be the vector of size $|\mathcal{V}|$ in which every row represents a different word used in the network. Let w_i be the word in place i of vector $\mathcal{V}$. It is possible to represent post p_j as a sequence of a set of S_j words of $\mathcal{V}$, with $S_j = |p_j|$, where $j \in \{1, \ldots, |\mathcal{P}|\}$ and $\mathcal{P}$ corresponds to the set of posts in the network. A "corpus"is defined as a collection of posts $\mathcal{C} = \{p_1, \ldots, p_N\}$. In terms of composition we can define the matrix $\mathcal{W}$ of size $|\mathcal{P}| \times |\mathcal{V}|$ where each element of the matrix is defined as

$$w_{i,j} = \text{ number of times } w_i \text{ appears in } p_j \tag{3.1}$$

Then $\sum_{i=1}^{|\mathcal{V}|} w_{i,j} = S_j$. Likewise, we can define $\sum_{j=1}^{|\mathcal{P}|} w_{i,j} = T_i$ which represents the total number of appearances the term w_i makes in a corpus.
salton1975vector A corpus can be represented by the product between the frequency of a term in the corpus and the logarithm of the reciprocal of the frequency of the document that contain the word (TF-IDF) [2]. We can define the tf-idf matrix that represents the corpus as $\mathcal{M}$ of size $|\mathcal{P}| \times |\mathcal{V}|$ where each $m_{i,j}$ is determined as

$$m_{i,j} = \frac{w_{i,j}}{T_i} \times \log\left[\frac{|\mathcal{P}|}{1 + n_i}\right] \tag{3.2}$$

Where n_i is the number of posts that belong to the corpus in which the word w_i appears in. The IDF term presented in 3.2 corresponds to a usual correction with respect to the original IDF term $\log\left[\frac{|\mathcal{P}|}{n_i}\right]$ because, after the cleaning and selection of data, some posts may not contain words, causing this term to be undefined.

3.1.3 Latent Dirichlet Allocation

Next, the process of content reduction will be presented through the use of a topic model, based on the work done by Ríos et al. [64], Álvarez [33] and L'Hullier [35]. A topic model can be considered as a probabilistic model that relates documents and words through variables which represent the main topics inferred from the text itself. In this context, a document can be considered as a mixture of topics, represented by probability distributions which can generate the words in a document given these topics. The inferring process of the latent variables, or topics, is the key component of this model, whose main objective is to learn from text data the distribution of the underlying topics in a given corpus of text documents.

Latent Dirichlet Allocation (LDA) [17, 28] is a Bayesian model where the latent topics in the documents are inferred through the estimation of distributions over a set of training data. The purpose is that each topic be modeled as a probability distribution over a set of words represented by the vocabulary ($w \in \mathcal{V}$), and every document as a probability distribution over a set of topics ($\mathcal{T}$). The sampling of these distributions is done with multinomial Dirichlet distributions.

The process is carried out in an automated way and it is only necessary to label each topic discovered by the algorithm with the help of experts from the community.

Using the definitions from section 3.1.2 and considering that a message contained in a post can be represented as a sequence of S words defined as $\mathbf{w} = (w^1, ..., w^S)$, where w^s represents the s^{th} word in the post, we can describe the generative process for the LDA model devised by Blei [17].

For each post of the corpus:

1. Choose a number S of multinomials ($S \sim Poisson(\xi)$) that will represent the number of words in the post.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each $w^s \in (w)$:
   (a) Choose a topic $z_s \sim Multinomial(\theta)$.
   (b) Choose a word w^s of $p(w^s|z_s, \beta)$, which is a multinomial conditional probability over the topic z_s.

For LDA, given the smoothing parameters β and α, and a joint distribution of a topic mixture θ, the idea is to determine the probability distribution to generate from a set of topics $\mathcal{T}$, a message composed by a set of S words $\mathbf{w}$,

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p(w^s|z_s, \beta) \tag{3.3}$$

Where $p(z_s|\theta)$ can be represented by the random variable θ_i, such that topic z_s is presented in document i ($z_s^i = 1$). A final expression can be deduced by integrating (3.3) over the random variable θ and summing over topics $z \in \mathcal{T}$. Given this, the marginal distribution of a message can be defined as shown in (3.4):

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{s=1}^{S} \sum_{z_s \in \mathcal{T}} p(z_s|\theta) p(w^s|z_s, \beta) \right) d\theta \tag{3.4}$$

The final goal of LDA is to estimate previously described distributions to build a generative model for a given corpus of messages. There are a large number of methods developed to perform the inference on this probability distribution, such as, variational expectation-maximization [17], or a discrete variational approximation of the equation 3.4 empirically by Xing [27] and through a Gibbs sampling (Monte Carlo model based on Markov chains) [19] which has been efficiently implemented and applied by Phang and Nguyen [29].

According to Ríos's methodology [64], the next step for the formation of the matrix [Topics× Posts]. This matrix will deliver a reduction in the content of the posts. We will use the method developed by Phang and Nguyen [29] which requires a series of input parameters, such as number of iterations, hyper parameters α, β, number of required topics $\mathcal{T}$, number of words n per topic that should be saved, posts of the corpus, among others. The method will return as output the distributions of the words over the topics, the distribution of these over the document, k topics with the n most important words that represent the topic and their corresponding probabilities of belonging to it, among other things. In particular, with the k topics and their n representative words or terms you can form vectors of size $|\mathcal{V}|$ by completing with zeros in the $|\mathcal{V}|-n$ non-representative words. With these vectors the semantic matrix (SM) [Topics×Terms] will be formed, then this matrix will be operated with the tf-idf transposed matrix $\mathcal{M}^t$ defined by the equation (3.2) that represents the participation of each term in each post, thus obtaining the matrix [Topics×Posts]. This matrix allows to know the composition

of a post in terms of the different topics found and allows us to obtain all of the post's text vector representations (ρ_p) where p is a post in the set of Posts.

3.2 ETL for Model Input Data

In forums, the social network graph is not explicitly defined as in other communities (Facebook, Twitter, etc.). Thus, we first must define a network topology to use. With this in mind, the most usual (and most direct) network representation used is one were every node represents a user of the network and a link is added between nodes to represent a relationship or interaction between the users they represent. Nevertheless, there are many feasible ways of representing the network that way. Some of these ways of defining the links in the network in web forums are shown in Fig. 3.2

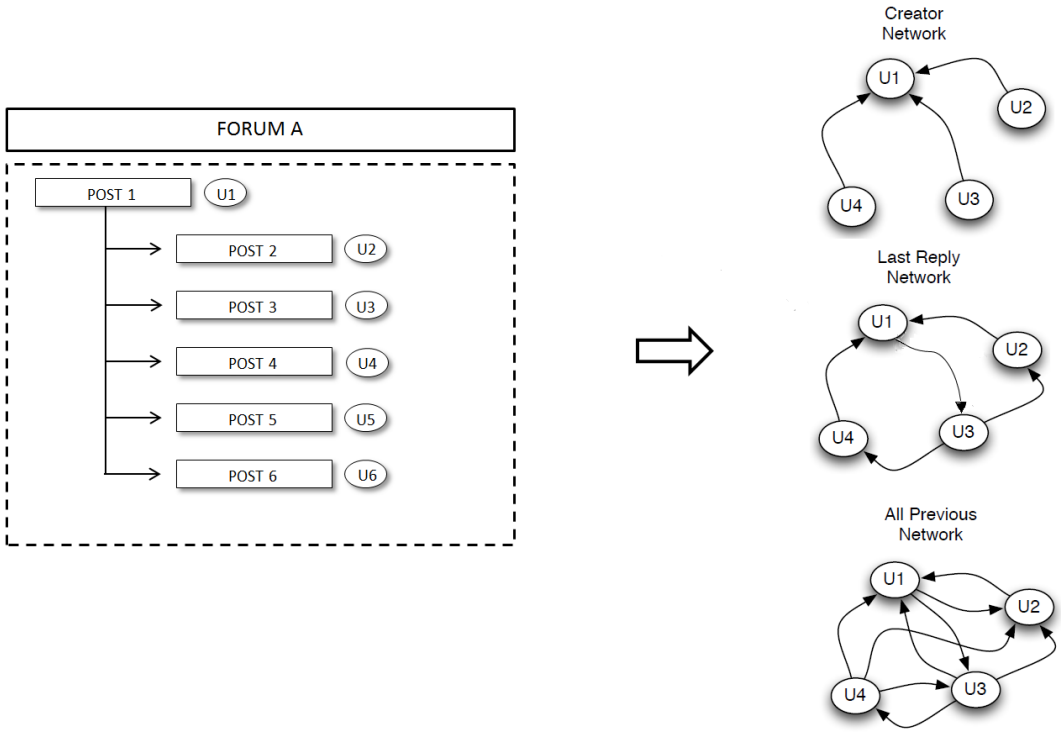


Figure 3.2: Possible Network Representations For Web Forums.

However, in accordance to our goal of making a content-centered model, we

propose a new network topology that puts emphasis on content and how users interact with it. This topology, as can be appreciated in Fig. 3.3, consists in distinguishing between four types of nodes, namely, Forum node, Sub-Forum nodes, Thread nodes and User nodes. These nodes follow a hierarchy where a type of node can only form a link with a node of a type belonging to a layer directly above them. We will mostly focus on the links formed between User nodes and Thread nodes. Note that users now interact (form links) directly with the conversations (Threads) that catch their interest which represents accurately what happens in web forums.

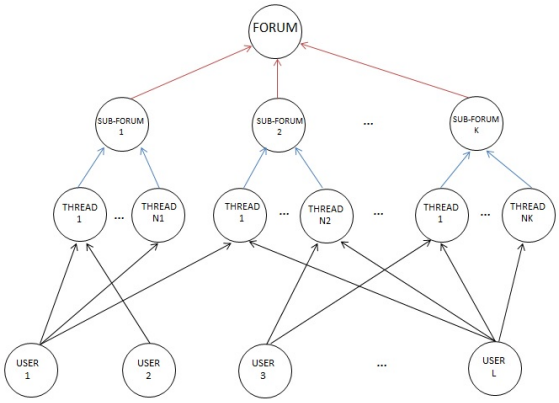


Figure 3.3: Proposed Topology For Web Forums.

Moreover, we can always derive the network representation were users interact with each other as if we have chosen the Creator-oriented Network representation. This can be easily seen in Fig. 3.4 were we show the equivalence (in terms of links formed) between the usual topology and the topology we proposed. Furthermore, if any of the usual ways of representing the network were chosen then if the number of active users in the network is n then the number of possible arcs would be given by

$$\text{number of possible arcs in the network} = n(n-1) \tag{3.5}$$

We can reduce the number of possibilities by adopting our proposed network topology. By representing the networks with our proposed topology, we have that if the number of active users in the network is n and the number of active threads is m then the number of possible arcs would be given by

$$\text{number of possible arcs in the network} = nm \tag{3.6}$$

where usually $m \ll n$ (if this is not the case it is always possible to shorten the time window considered as a period and thus make the former inequality true). This is of particular importance when making the model choose the correct link to be formed.

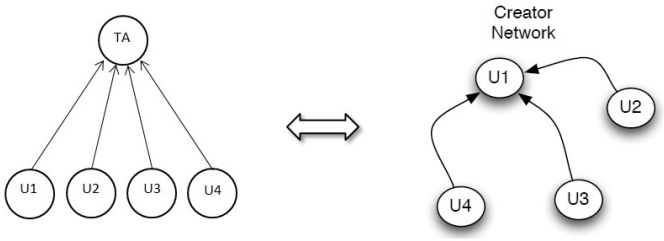


Figure 3.4: Equivalent Heterogeneous Network Representation.

As a final remark, using the topology we propose helps to incorporate additional aspects of the network, as shown in Fig. 3.3, particularly regarding the structure of the Forum. In this work, we extracted the information directly from the forum structure for the first three layers, i.e., we will only try to predict links between User nodes and Thread nodes.

After establishing the network representation to be used in this work we had to face the issue of using the text content generated by user's posts in order to extract hints regarding which arcs are more likely to exist, i.e., which conversations are more likely to be appealing to which users. Having the text vector representation of each of the posts available in the data within the time frame established as a result of the preprocessing stage we face the problem of getting each user's text vector representation and thread's text vector representation. There are multiple alternatives for solving this problem but for this work we decided to use the following proposed approach mainly due to it outperforming other approaches. First, we subdivide the posts of the Forum into different groups according to the Sub-Forum they belong to. Next, we subdivide the time frame into periods and subsequently we subdivide the posts within each Sub-Forum into different groups according to the time period they belong. Then we extract the period's active users and threads by considering a user as active if he makes a post during the period and a thread as active if a user posted in the thread during the period. After that, we compute a thread's text vector representation for a period, ν_T^P, as the mean of all of the post's text vector representations, ρ_p, of the posts that belong to both the thread,

T, and the period, P, i.e.

$$\nu_T^P = \frac{1}{|T \cap P|} \sum_{p \in T \cap P} \rho_p \tag{3.7}$$

where p is a post. On the other hand, to compute a user's text vector representation we first need to make one more subdivision. We subdivide the posts made by a user during the period into different groups according to the thread they were posted in. It is important to notice that a user will have as many text vector representations for a period as there are subgroups of his posts during the aforementioned period. We can now compute a user's text vector representation for a period and subgroup of posts, μ_S^P, as the mean of all of the post's text vector representations, ρ_p, of the posts that belong to both the subgroup of posts, S, and the period, P, i.e.

$$\mu_{U,S}^P = \frac{1}{|U \cap S \cap P|} \sum_{p \in U \cap S \cap P} \rho_p \tag{3.8}$$

where p is a post. Now that we have the text vector representations for both users and threads sorted out we must apply the process shown in Fig. 3.5 to be able to obtain something useful for the model.

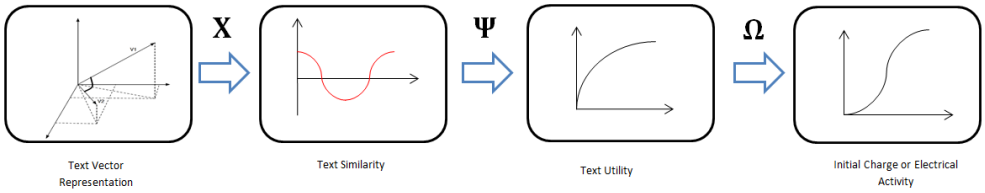


Figure 3.5: Transformations applied to get model input.

First, we need need to define a function χ that gives us a measure of how related or distant are two text vector representations to each other. To compute this, we make use of cosine similarity measure which gives us the cosine of the angle formed between two text vector representations in a k-topic space. Thus, if we have two text vector representations μ_1 and μ_2 the similarity between them is given by

$$\chi(\mu_1, \mu_2) = similarity(\mu_1, \mu_2) = cos(\theta) = \frac{\mu_1 \cdot \mu_2}{|\mu_1||\mu_2|} \tag{3.9}$$

where θ is the angle between μ_1 and μ_2. Now we face a different problem which is to define a function Ψ to get from text vector representation similarity to user utility.

The answer to this question is not a straight forward one mainly due to that it is not clear the exact way a user reacts to similar texts. This problem has not been regarded in previous investigations at least when formulating diffusion models. We propose, as a way to bridge this obstacle, to make a transformation via a function that allows to extract in a better way the characteristics or differences between alternatives having in mind that this utility will serve as input for a logit model. An interesting question to answer is how we can determine the optimal functional form to transform text similarity to user utility, but this question falls beyond the scope of this research. So, having stated the problem we can conceptualize it as determining a function Ψ_1 that takes text similarity, x, and gives user utility as output. We propose the following functional form

$$\Psi_1(x) = \frac{1}{1-x} \tag{3.10}$$

because it maximizes the impact generated by differences in input to logit probabilities. Nevertheless, taking into consideration that the former function can have a very wide range of values we can attempt to control this range and constrain it to the $[0, a]$ interval by introducing the following transformation to the utility by defining the function Ψ_2 such that

$$\Psi_2(u) = a\frac{u}{w} \tag{3.11}$$

where $w = \max_{j \in Threads} u_j$ and a is a parameter that determines the length of the interval in which the values of the transformed utility will vary. This makes a a fundamental parameter in terms of importance because depending on the value chosen it can constrain the dispersion of the logit probabilities or conversely permit to much variability. Taking this information into consideration we can consider a as a measure of stochasticity allowed into the initial probabilities. For the model a represents the sensibility of the network users to how similar a conversation's topic is to the user's generated text content in terms of satisfaction extracted from that conversation.

Before moving on to the next step we chose to reduce the number of alternatives a user considers during his decision-making process. To accomplish this, for each choice the user makes, we sorted the alternatives (active threads) according to their reported utility (V_j) to the user in a decreasing fashion and decided to keep only the top k alternatives. This reduction of alternatives a user considers is based on research about working memory and attention span [1].

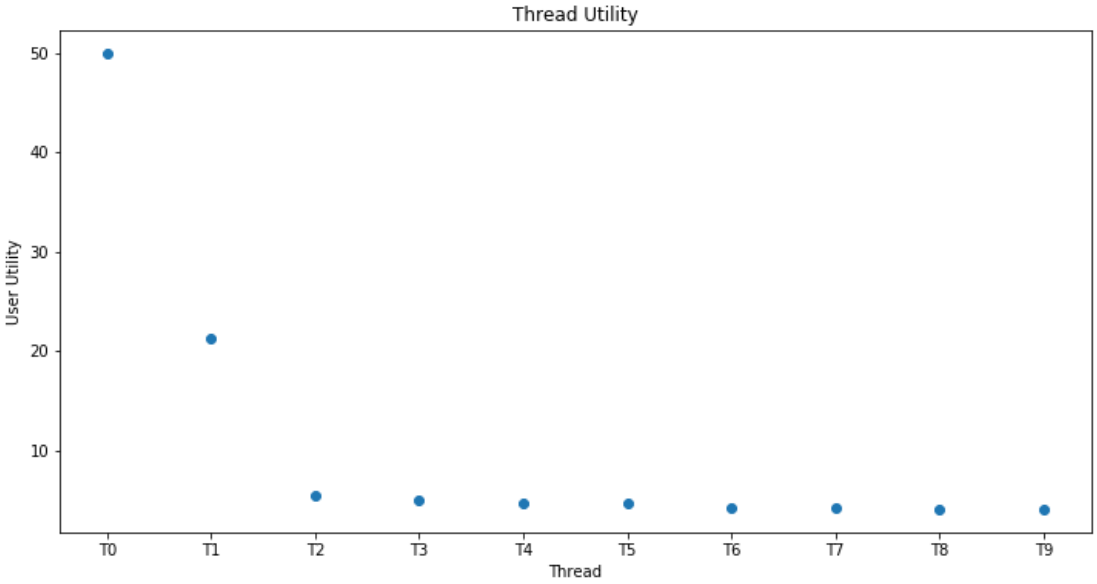


Figure 3.6: Example 1 of thread utility

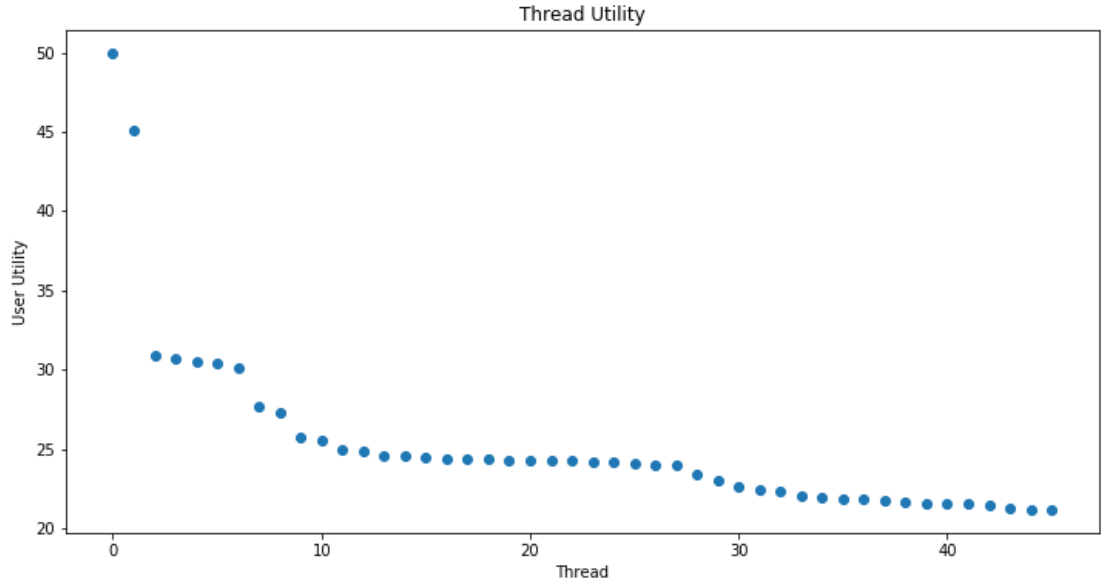


Figure 3.7: Example 2 of thread utility

In Fig. 3.6 and 3.7 we can see 2 examples of the utility a user finds in all of the threads that are active at that time. We can notice that only a few threads

are of great interest to the user and most threads are stacked at the tail of the figure adding noise to the decision. Thus, keeping only the k alternatives with most utility helps to reduce the noise we add to the model.

As a final step we need to define a function Ω such that we get something resembling the initial electrical activity of the neural region associated with a certain decision. For this purpose, we decided to make use of random utility theory in the way that I_i is proportional to the likelihood of choosing alternative i as we can appreciate in

$$\Omega(\mathbf{V}, i) = I_i = \frac{e^{V_i}}{\sum_{j \in Threads} e^{V_j}} \tag{3.12}$$

3.3 Neurophysiological Model Setup

3.3.1 Leaky Competing Accumulator

This proposal is based on decision-making Psychological models building on the work done in [48], and we will consider that OSN users correspond to cognitive agents in the sense explained in this section. Psychology considers cognitive (from mental operation) and perceptual (from sensory devices) processes, as well as motor actions. Memory retrieval/storage, language production/interpretation, attention, problem solving, and decision-making are example of some cognitive processes. The LCA model was presented by Usher and McClelland [15] in 2001 as a diffusion model for decision making. Their work considers a theoretical unification of concepts from cognitive-perceptual processes and the underlying neurophysiology. The model describes the stochastic evolution of average electrical neural activities $\{X_i\}$ of certain brain regions $\{i\}$. Each i labels a possible decision that a cognitive agent is going to decide whether to take or not. The process evolves according to (3.13), starting with $X_i(t = 0) \sim 0^+$, and stopping at time $t = T(i^*)$. The stopping condition is fired when a neural activity value first reaches a given threshold $X_{i^*} = Z$, in which case the decision taken is i^*.

$$dX_i = \left[I_i - \sum_j \omega_{ij} X_j \right] dt + \sigma_i dW_i, \quad i = 1, \ldots, M \tag{3.13}$$

Other terms in (3.13) correspond to exogenous parameters. I_i is an input value in favor of alternative i that is accumulated from other devices such as the visual

cortex, and serves as an input to the LCA process. Those values are supposed to be constrained as $I_i \geq 0$ under neurophysiology reasoning. External input values are accumulated in the variable X_i in favor of alternative i. The ω parameter is represented by two values, as shown in (3.14).

$$\omega_{ij} = \begin{cases} \kappa & i = j \\ \lambda & i \neq j \end{cases} \tag{3.14}$$

The κ parameter from (3.14) takes into account the decay [5]. Lateral inhibition between accumulator units is controlled by the λ parameter and considers equal effect for all units. The accumulated values are considered biological values, such as neural activity (rate of spikes), which are then restricted to being positive ($X > 0$). According to [22] perceptual decision making consists of three brain processing stages. First, a representation of sensory evidence is carried out after recollection from the sensory system. Second, integration of the available sensory information is performed over time on an appropriate buffer. Finally, either a comparison is made between evidence in favor of the involved decisions, or a threshold triggers the choice in some cases.

3.3.2 Customization of LCA

We make two modifications to the LCA model exposed in section 3.13. First, we make a segmentation of the users according to their level of involvement in the forum. This classification into types of users is obtained from forum's community experts. We believe it is reasonable to assume their decision-making process is affected in a different way by content given that they behave in differently in their participation in the community.

Furthermore, we modify the LCA model including a term that accounts for habit formation by reinforcing the probability of choosing a alternative that has been chosen before. This is done like we show in (3.15)

$$X_i(t = 0) = 0 + (1 + \gamma)^{\text{number of successes}} - 1 \tag{3.15}$$

where $\gamma \geq 0$. It is important to understand that this is a first approximation to incorporating habit formation into the model. According to habit formation literature, increase in automaticity by repetition follows an asymptotic curve [37] but this is in the long run. Taking into consideration that we wish to incorporate the effects of habit formation during an initial stage then our approximation is acceptable.

We can see the pseudocode structure of the proposed model in Algorithm 1.

Algorithm 1 LCA Link Prediction Model

Require: $\{\mathcal{U}, \mathcal{T}, \mathcal{P}, \mathcal{A}^{real}\}$

Set $\{\lambda_k\}_{k\in C}$,$\{\kappa_k\}_{k\in C}$,$\{\beta_k\}_{k\in C}$,z,dt,γ,a
for each $t \in \mathcal{T}$ **do**
  $T_t^0 \leftarrow \sum_{p\in p(t)} \frac{\Lambda(\theta(p))}{|p(t)|}$
end for
for each $u \in \mathcal{U}$ **do**
  $u_{threads} \leftarrow |\mathcal{T}(u)|$
  for $i \leq u_{threads}$ **do**
    $U_{u,i}^0 \leftarrow \sum_{p\in p(u,t(i))} \frac{\Lambda(\theta(p))}{|p(u,t(i)|}$
  end for
end for
for each $u \in \mathcal{U}$ **do**
  for $i \leq u_{threads}$ **do**
    for each $t \in \mathcal{T}$ **do**
      $V_{u,i,t}^0 \leftarrow \frac{1}{1-COS(U_{u,i}^0, T_t^0)}$
    end for
  end for
end for
for each $u \in \mathcal{U}$ **do**
  for $i \leq u_{threads}$ **do**
    for each $t \in \mathcal{T}$ **do**
      $V_{u,i,t} \leftarrow a\frac{V_{u,i,t}^0}{\max_{j\in\mathcal{T}} V_{u,i,j}^0}$
    end for
  end for
end for
for each $u \in \mathcal{U}$ **do**
  $u_{class} \leftarrow C(u)$
  for $i \leq u_{threads}$ **do**
    for each $t \in \mathcal{T}$ **do**
      $I_{u,i,t} \leftarrow \beta_{u_{class}} \frac{e^{V_{u,i,t}}}{\sum_{j\in\mathcal{T}} e^{V_{u,i,j}^0}}$
      $\sigma_{u,i,t} \leftarrow \sqrt{1.5 I_{u,i,t}}$
    end for
  end for
end for
for each $u \in \mathcal{U}$ **do**
  for $i \leq u_{threads}$ **do**
    for each $t \in \mathcal{T}$ **do**
      $a_{u,i} \leftarrow$ sample $X^{LCA}(\ \lambda_{u_{class}}, \kappa_{u_{class}}, I_{u,i,t}, \sigma_{u,i,t}, z, \gamma)$
      $\mathcal{A}^{sim} \leftarrow \mathcal{A}^{sim} \cup a_{u,i}$
    end for
  end for
end for
return $\mathcal{A}^{sim}$

3.4 Model Fitting

Finally, we implemented a genetic algorithm heuristic in order to obtain approximately optimal parameters for the model based in the implementation done in [69] where we used the linear-ranking algorithm of Baker [63] as the fitness function which indicates how well the current population fits the objective function. For reproduction in the algorithm we used roulette wheel selection [45] as selection mechanism and single point [7, 8] as the crossover routine. Finally, to perform the mutation operation we used real-value mutation [6].

Chapter 4

Experiments, Results and Evaluation

4.1 Plexilandia's Data

Plexilandia is an OSN formed by a group of people who have met towards the building of music effects, amplifiers and audio equipment ("Do it yourself" style). In the beginning it was born as a social network for sharing common experiences in the construction of plexies[1]. Today, plexilandia counts more than 2500 members in more than 15 years of existence. All these years they have been sharing and discussing their knowledge about building their own plexies and effects. Besides, there are other related topics such as luthier, professional audio and buying/selling parts.

Although, they have a basic social network information web page, most of their member's interactions are produced on the discussion forum. During nine years of its life, this OSN has undergone a great sustained growth in members and their contributions. In Table 4.1 we can see the number of posts for each of these 9 years including the total number of posts. We must note that for this work we only used the data of year 2013 and 2014, and Sub-Forums 2-6 as can be seen later in Fig.4.1.

[1] "Plexi" is the nickname given to Marshall amp heads model 1959 that have the clear perspex (a.k.a plexiglass) fascia to the control panel with a gold backing sheet showing through as opposed to the metal plates of the later models.

Table 4.1: Plexilandia Activity

Forum	***2006***	***2007***	***2008***	***2009***	***2010***	***2011***	***2012***	***2013***	***2014***	***TOTAL***
Aplifiers (2)[2]	392	2165	2884	3940	3444	3361	2398	1252	985	20822
Effects (3)	184	1432	3362	3718	4268	5995	4738	2317	1331	27345
Luthier (4)	34	388	849	1373	1340	2140	926	699	633	8382
General (5)	76	403	855	1200	2880	5472	3737	1655	1295	17573
Pro Audio (6)	—	—	—	—	—	342	624	396	219	1579
Synthesizers (7)	—	—	—	—	—	—	—	104	92	196
TOTAL	**686**	**4388**	**7950**	**10231**	**11932**	**17310**	**12423**	**6423**	**4555**	**75898**

In addition, Social network administrators provided us with a list of 64 key-members. As noted in [71] there are many definitions of what a key-member is, e.g. the users who participate most, the ones that answer other's questions, etc. but it is clear that they play a primordial role in keeping the network alive. Three groups of importance were established as follows:

- Experts Type A: which are the most important key-members. There are 34 members for year 2013 based on administrators' criteria.
- Experts Type B: which are the most important in a lesser degree than A type key-members, however, they are also key-members. These are 21 for the same period.
- Experts Type C: finally, C type key-members are those that are historic key-members, since they have been involved in the social network since its origins, but they are not continuously participating. In this class there are about 11 members for year 2013.
- Experts Type X: these class is all members of the social network which are not key-members. They don't belong to the social network core and usually they ask questions rather than publish answers, or tutorials.

It is important to remark that this information was given during 2013 and 2014 [71] so the probability of them forgetting key members was minimized. This constitutes the main reason we decided to use only the data of these years for our experiments. We used this classification of users as a segmentation of the population regarding behavior as noted in section 3.3.2.

4.2 Experimental Setup

From the dataset obtained from Plexilandia's administrators we extracted the following information for each of the post made between January 2013 and January 2014: User ID, Post ID, Thread ID, Sub-Forum ID, Post text content and Post time. We applied all of the data preprocessing described in section 3.1 to Posts text content to obtain text vector representation for each post. Additionally, we created a feature named User type in accordance to the information obtained about the network's key-members, distinguishing between 4 types of user as explained in section 4.1.

Consequently, with our proposed network topology we divided the dataset into sub-forums. By exploring the number of posts during different time periods (1 week, 2 weeks, 1 month, 2 months, 4 months) and the behavior of the threads during that time we decided to choose a time period size of 1 month obtaining a total of 13 time periods. In Fig. 4.1 we show the way we partitioned the data and how we chose to conduct the experiments, using the first month of 2013 (January) as calibration data for the model and the rest of the months as testing data. After

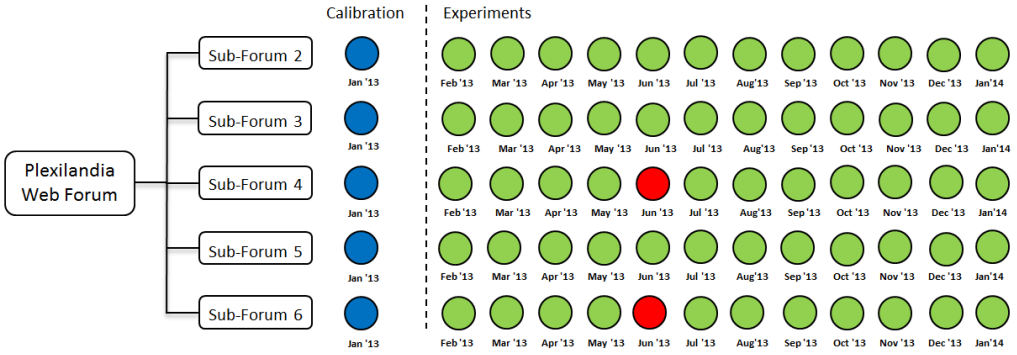


Figure 4.1: Experimental Setup

the separation was made we were able to compute the number of active users, active threads and posts made during each of these 13 months for each of the Sub-Forums as we show in Tables 4.2, 4.3 and 4.4.

Table 4.2: Active Users, Active Threads and Posts made in Sub-Forums (a) 2 and (b) 3

Month	Users	Threads	Posts
1	45	25	103
2	19	10	51
3	35	20	83
4	38	27	133
5	32	22	55
6	33	22	94
7	26	14	57
8	38	24	127
9	35	17	94
10	35	23	110
11	38	22	121
12	31	19	94
13	27	14	59
Total	**168**	**221**	**1181**

(a) Statistics of Sub-Forum 2

Month	Users	Threads	Posts
1	49	43	145
2	46	29	169
3	51	46	252
4	53	43	196
5	51	44	184
6	52	38	208
7	49	32	173
8	42	37	171
9	43	33	174
10	44	29	138
11	43	24	124
12	49	38	156
13	31	30	102
Total	**174**	**351**	**2192**

(b) Statistics of Sub-Forum 3

Table 4.3: Active Users, Active Threads and Posts made in Sub-Forums (a) 4 and (b) 5

Month	Users	Threads	Posts
1	32	40	115
2	25	8	81
3	20	13	60
4	22	15	50
5	12	8	23
6	5	3	7
7	19	10	46
8	21	17	57
9	19	10	52
10	20	9	30
11	22	9	72
12	12	8	33
13	28	17	104
Total	**96**	**134**	**730**

(a) Statistics of Sub-Forum 4

Month	Users	Threads	Posts
1	60	37	164
2	47	27	131
3	58	30	182
4	36	23	84
5	55	28	145
6	53	36	202
7	55	35	176
8	45	29	116
9	25	19	72
10	34	25	66
11	25	13	41
12	42	25	105
13	38	24	98
Total	**171**	**282**	**1582**

(b) Statistics of Sub-Forum 5

Table 4.4: Active Users, Active Threads and Posts made in Sub-Forum 6

Month	**Users**	**Threads**	**Posts**
1	14	11	49
2	7	5	13
3	16	6	33
4	6	5	13
5	11	9	30
6	11	5	13
7	10	7	52
8	9	3	13
9	11	7	41
10	15	5	27
11	8	5	37
12	15	6	36
13	11	6	27
Total	**50**	**47**	**384**

We proceed, as described in section 3.3.2, to assign text vector representation for each month to every active thread by using the mean vector of all posts belonging to that thread during the month that is being tested. As for users, we assign m text vector representations by grouping posts according to the thread to which they belong and computing the mean vector of the group of posts. This way we also recover the number of threads (m) with whom a user forms a link in our proposed network representation.

In order to evaluate the quality of our proposal, we decided to use the following evaluation framework. We will compute 4 measures to assess the performance of the model. Namely Recall, Accuracy, Precision and F measure [65]. A description of these metrics is given below.

Recall gives a measure of the probability of detection of the model and is defined as:

$$Recall = \frac{\text{Number of true positive links}}{\text{Number of real links}} \tag{4.1}$$

Accuracy gives a measure of the trueness of the results in the sense that it describes systematic observational errors or statistical bias in the model. Accuracy is defined

as:

$$Accuracy = \frac{\text{Number of true positive links} + \text{Number of true negative links}}{\text{Number of possible links}} \tag{4.2}$$

Precision gives a measure of statistical variability, or in other words it describes the random observational error of the model. It is defined as:

$$Precision = \frac{\text{Number of true positive links}}{\text{Number of predicted links}} \tag{4.3}$$

F measure or F_1 score combines the measures of precision and recall obtaining an alternative measure of the model's accuracy and is defined as:

$$\text{F measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{4.4}$$

4.3 Experimental Results

We run the GA to obtain the following optimized parameters for the LCA model

Table 4.5: Calibrated values of (a) β and (b) κ

Sub-Forum	$\beta_{\mathbf{A}}$	$\beta_{\mathbf{B}}$	$\beta_{\mathbf{C}}$	$\beta_{\mathbf{X}}$
2	0.863	0.148	0.511	0.553
3	0.584	0.906	0.389	0.029
4	0.586	0.833	0.352	0.476
5	0.628	0.184	0.000	0.429
6	0.516	0.126	0.490	0.595

(a) β calibrated values

Sub-Forum	$\kappa_{\mathbf{A}}$	$\kappa_{\mathbf{B}}$	$\kappa_{\mathbf{C}}$	$\kappa_{\mathbf{X}}$
2	0.174	0.055	0.070	0.965
3	0.684	0.340	0.217	0.588
4	0.642	0.389	0.866	0.981
5	0.707	0.733	0.047	0.623
6	0.287	0.692	0.087	0.401

(b) κ calibrated values

Table 4.6: λ calibrated values

Sub-Forum	$\lambda_{\mathbf{A}}$	$\lambda_{\mathbf{B}}$	$\lambda_{\mathbf{C}}$	$\lambda_{\mathbf{X}}$
2	0.491	0.137	0.399	0.189
3	0.146	0.951	0.189	0.949
4	0.639	0.478	0.107	0.245
5	0.0935	0.864	0.847	0.640
6	0.956	0.869	0.044	0.315

With these parameter values we used the model to simulate the networks for each Sub-Forum and for each month between Feb 2013 and Jan 2014. We then reconstructed the simulated and real network graphs according to our proposed network representation and we computed 4 metrics to evaluate performance with regards to links predicted. For each Sub-Forum we present the results obtained for these 4 metrics and 2 representative network images of the best and worst result in F measure for the time frame considered.

4.3.1 Sub-Forum 2

In Table 4.7 we show the results obtained for each of the metrics evaluated for Sub-Forum 2. As we can notice, the best result with regards to F measure is obtained in month 2 and the worst in month 4

Table 4.7: Results of Sub-Forum 2

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.724	0.916	0.724	0.724
3	0.525	0.924	0.554	0.539
4	0.435	0.910	0.457	0.446
5	0.511	0.939	0.523	0.517
6	0.473	0.924	0.5	0.486
7	0.643	0.920	0.659	0.651
8	0.527	0.928	0.557	0.542
9	0.566	0.928	0.6	0.583
10	0.556	0.937	0.603	0.579
11	0.485	0.917	0.493	0.489
12	0.526	0.910	0.536	0.531
13	0.667	0.934	0.684	0.675
Mean	**0.553**	**0.924**	**0.574**	**0.564**

Recall	Accuracy	Precision	F-measure
0.724	0.916	0.724	0.724

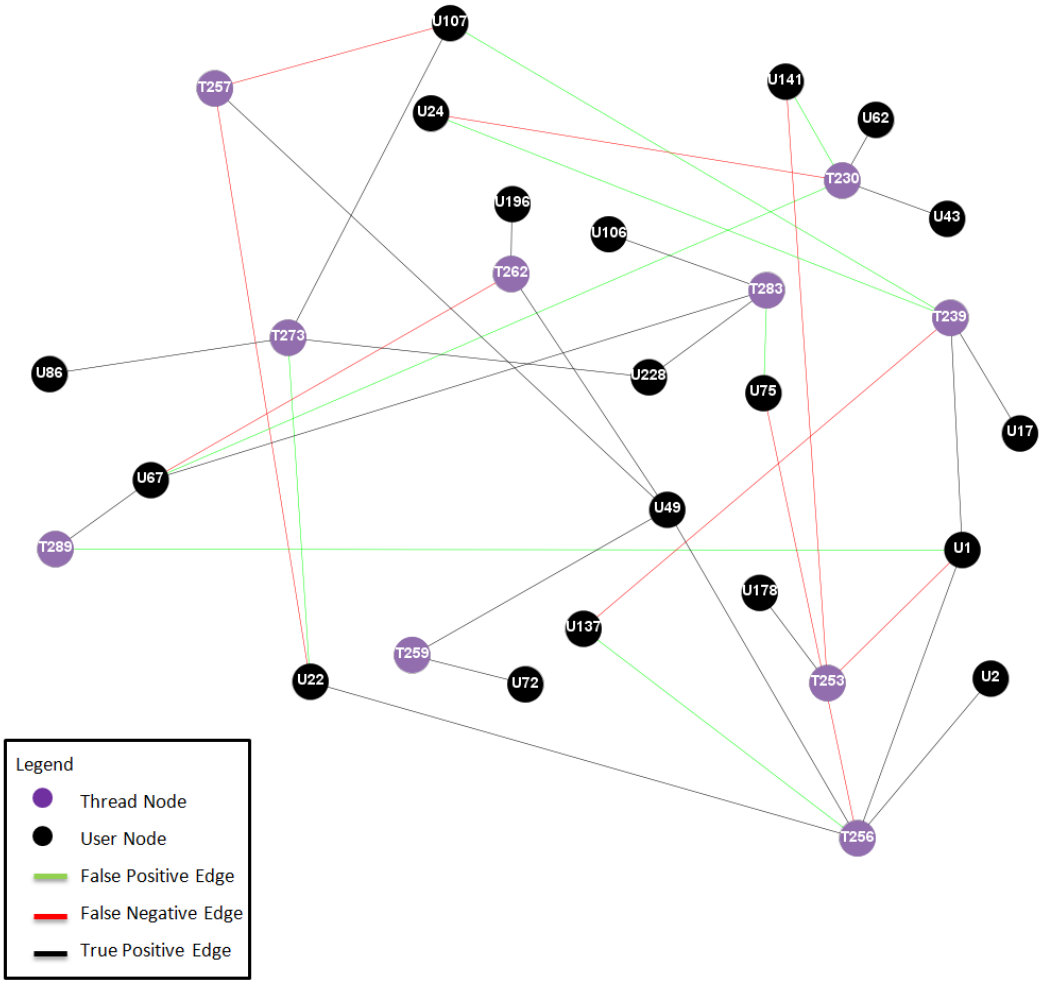


Figure 4.2: Network of Sub-Forum 2 for Month 2

Fig. 4.2 shows the network for sub-forum 2 for month 2 where thread nodes are shown in violet and user nodes are shown in black. Additionally, edges in black correspond to the edges the simulation predicted correctly, edges in green are edges the simulation predicted incorrectly and lastly edges in red correspond to the edges that the simulation failed to predict. We can notice that most of the network edges are black and that there is approximately the same amount of predicted links than real links.

Recall	Accuracy	Precision	F-measure
0.435	0.910	0.457	0.446

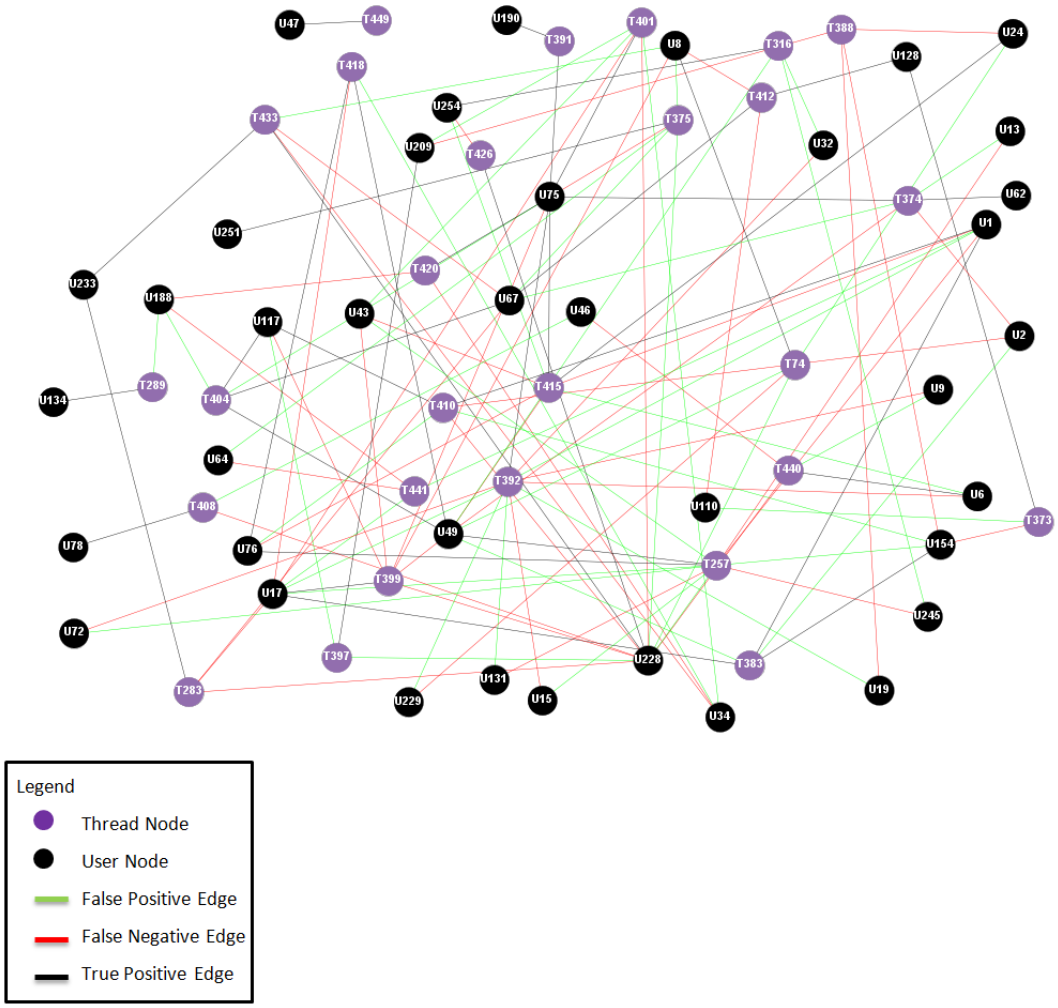


Figure 4.3: Network of Sub-Forum 2 for Month 4

Fig. 4.3 shows the network for sub-forum 2 for month 4 where thread nodes are shown in violet and user nodes are shown in black. Additionally, edges in black correspond to the edges the simulation predicted correctly, edges in green are edges the simulation predicted incorrectly and lastly edges in red correspond to the edges that the simulation failed to predict. We can notice that most of

the network edges are black and that there is approximately the same amount of predicted links than real links.

4.3.2 Sub-Forum 3

In Table 4.8 we show the results obtained for each of the metrics evaluated for Sub-Forum 3. As we can notice, the best result with regards to F measure is obtained in month 13 and the worst in month 11

Table 4.8: Results of Sub-Forum 3

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.432	0.909	0.453	0.442
3	0.453	0.929	0.477	0.465
4	0.496	0.943	0.515	0.506
5	0.431	0.939	0.445	0.438
6	0.496	0.939	0.508	0.502
7	0.429	0.925	0.441	0.435
8	0.455	0.929	0.495	0.474
9	0.440	0.911	0.451	0.445
10	0.556	0.939	0.568	0.562
11	0.410	0.917	0.445	0.427
12	0.471	0.943	0.485	0.478
13	0.632	0.951	0.672	0.652
Mean	**0.475**	**0.931**	**0.496**	**0.486**

Recall	Accuracy	Precision	F-measure
0.410	0.917	0.445	0.427

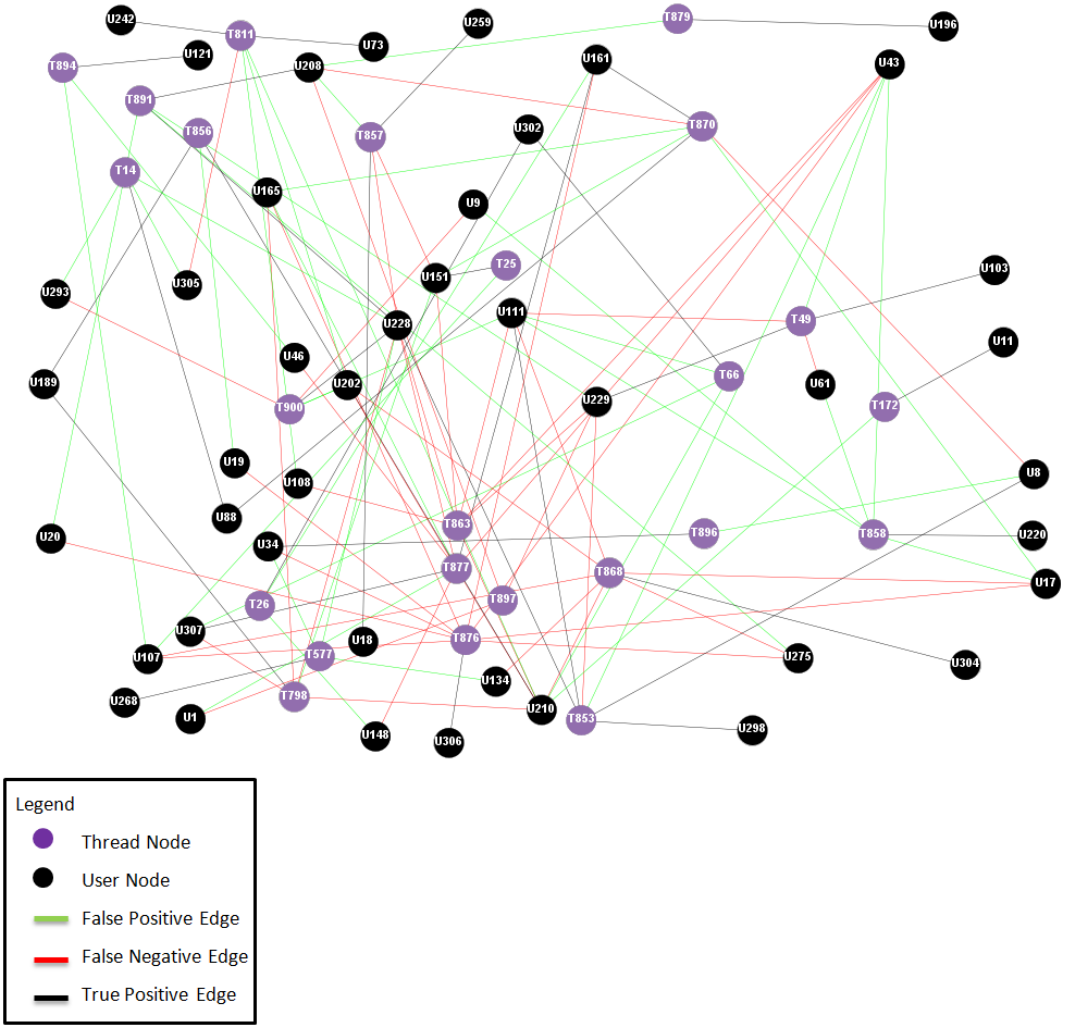


Figure 4.4: Network of Sub-Forum 3 for Month 11

Recall	Accuracy	Precision	F-measure
0.632	0.951	0.672	0.652

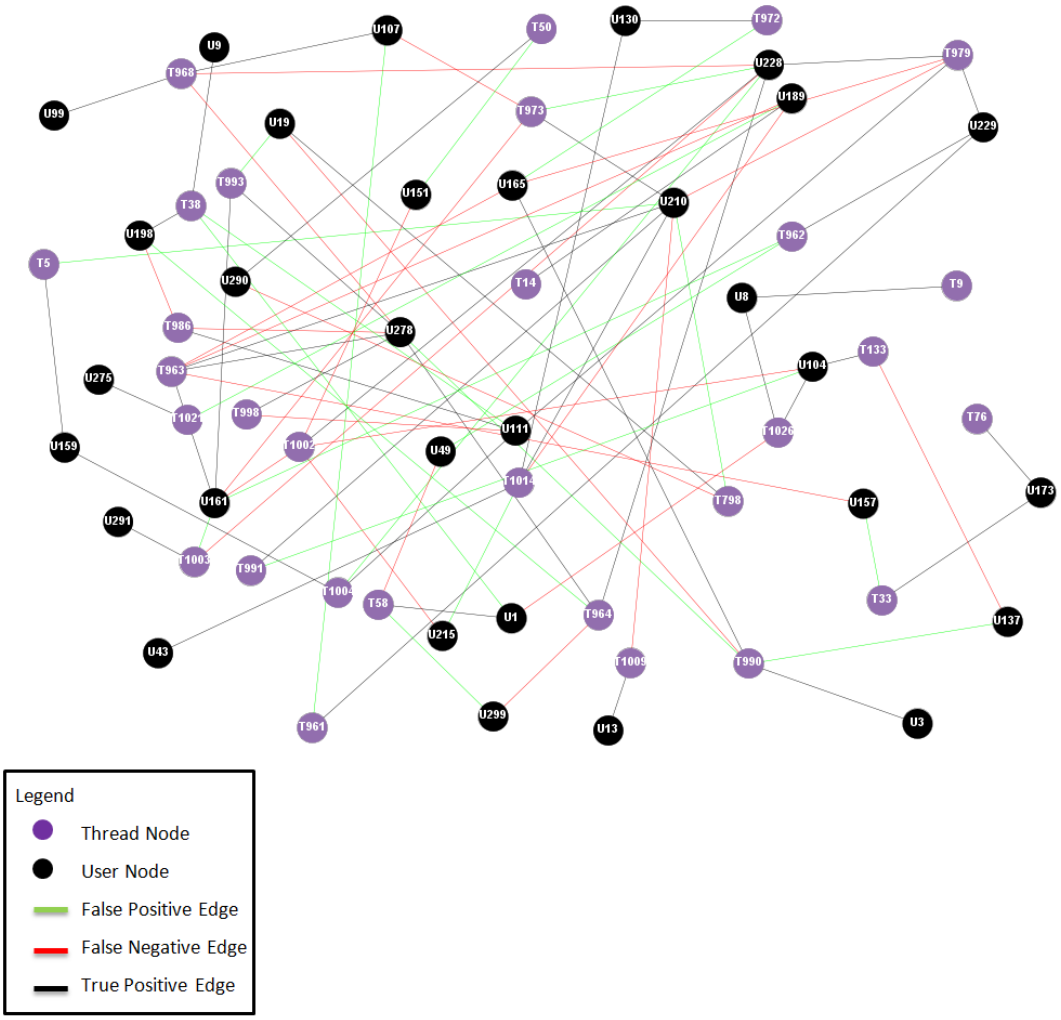


Figure 4.5: Network of Sub-Forum 3 for Month 13

4.3.3 Sub-Forum 4

In Table 4.9 we show the results obtained for each of the metrics evaluated for Sub-Forum 4. As we can notice, the best result with regards to F measure is obtained in month 5 and the worst in month 3. Due to the low number of posts, users and threads the results obtained for month 6 were discarded.

Table 4.9: Results of Sub-Forum 4

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.600	0.825	0.698	0.645
3	0.454	0.831	0.500	0.476
4	0.632	0.915	0.632	0.632
5	0.778	0.917	0.778	0.778
6	—	—	—	—
7	0.581	0.879	0.643	0.610
8	0.634	0.927	0.703	0.667
9	0.636	0.884	0.677	0.656
10	0.692	0.917	0.720	0.706
11	0.650	0.874	0.708	0.675
12	0.700	0.885	0.737	0.718
13	0.537	0.876	0.563	0.550
Mean	**0.627**	**0.885**	**0.669**	**0.647**

Recall	Accuracy	Precision	F-measure
0.454	0.831	0.5	0.476

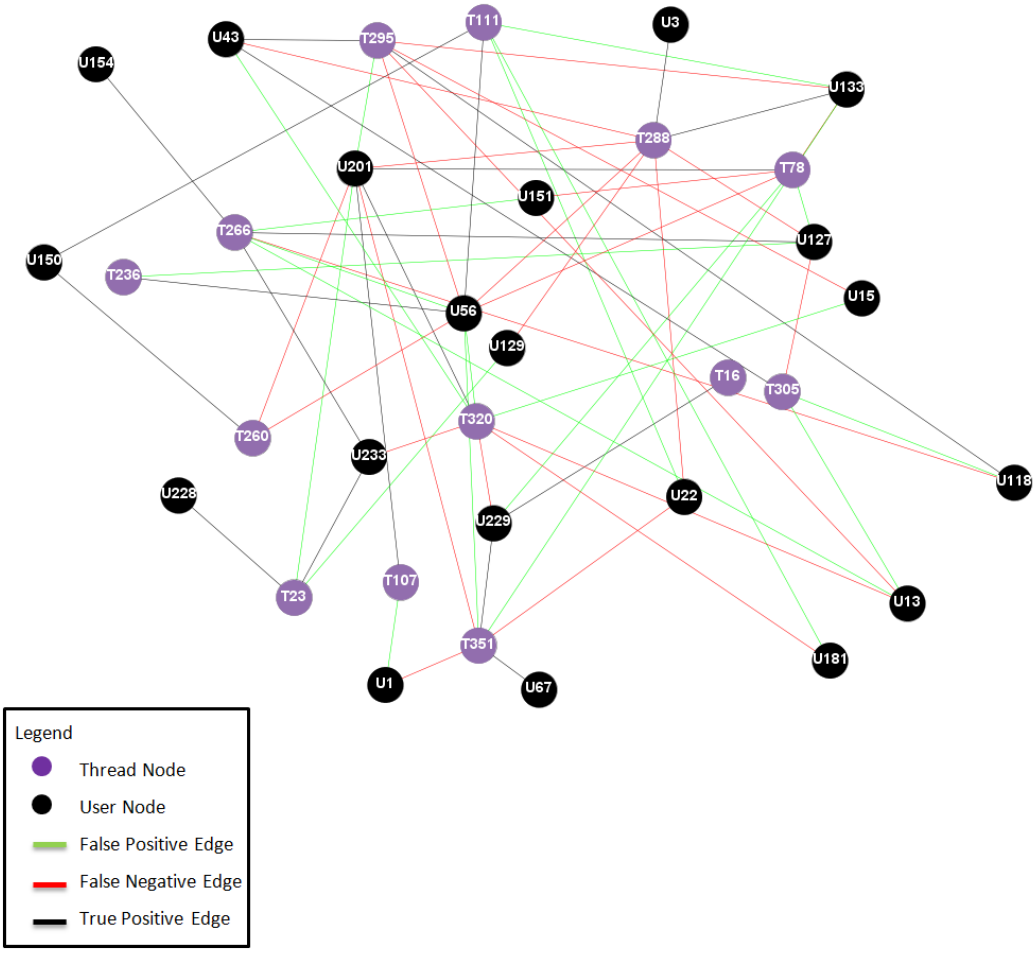


Figure 4.6: Network of Sub-Forum 4 for Month 3

Recall	Accuracy	Precision	F-measure
0.778	0.917	0.778	0.778

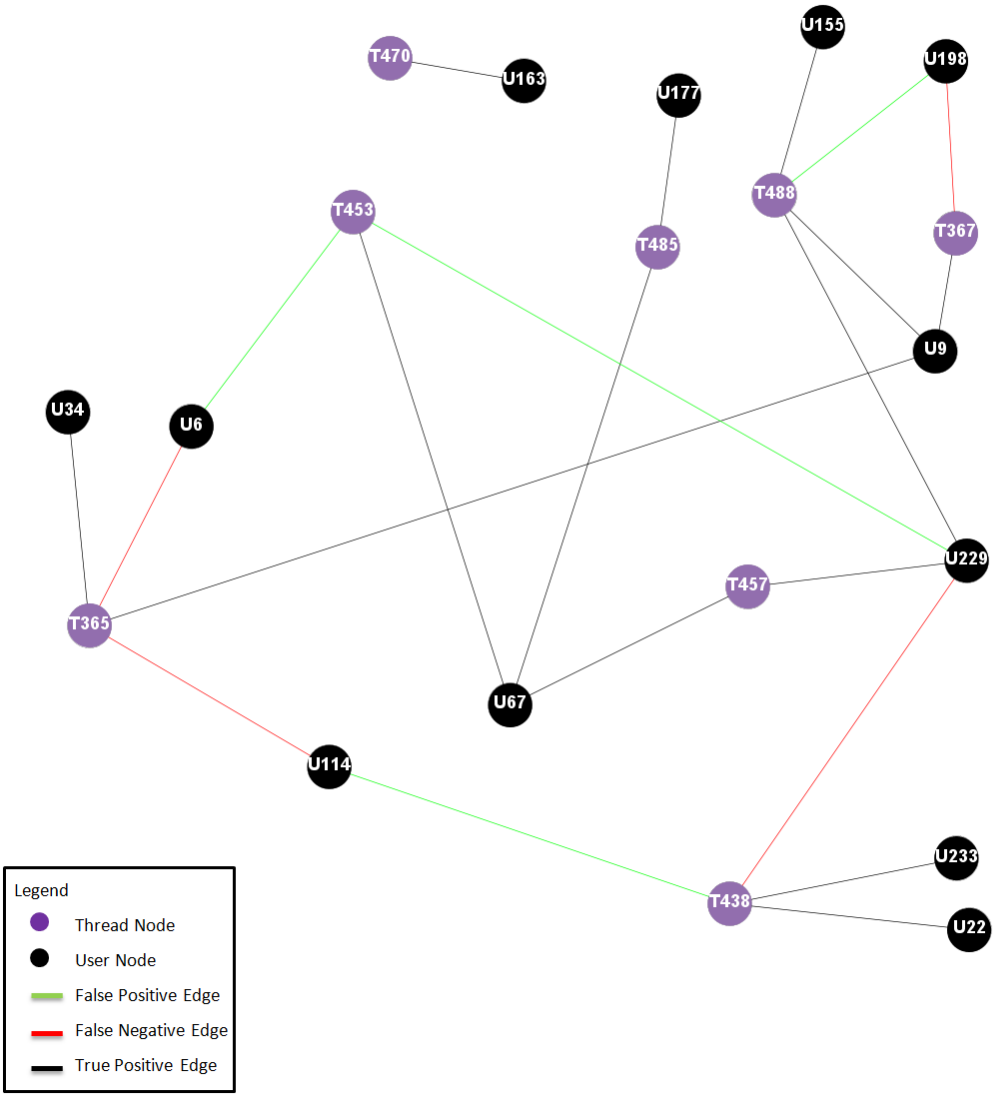


Figure 4.7: Network of Sub-Forum 4 for Month 5

4.3.4 Sub-Forum 5

In Table 4.10 we show the results obtained for each of the metrics evaluated for Sub-Forum 5. As we can notice, the best result with regards to F measure is obtained in month 9 and the worst in month 6.

Table 4.10: Results of Sub-Forum 5

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.474	0.939	0.500	0.487
3	0.431	0.931	0.448	0.439
4	0.557	0.936	0.567	0.562
5	0.453	0.928	0.475	0.464
6	0.377	0.919	0.402	0.389
7	0.470	0.938	0.478	0.474
8	0.457	0.926	0.483	0.470
9	0.674	0.939	0.689	0.681
10	0.615	0.955	0.640	0.627
11	0.647	0.926	0.647	0.647
12	0.507	0.933	0.521	0.514
13	0.443	0.907	0.461	0.452
Mean	**0.509**	**0.931**	**0.530**	**0.517**

Recall	Accuracy	Precision	F-measure
0.377	0.919	0.402	0.389

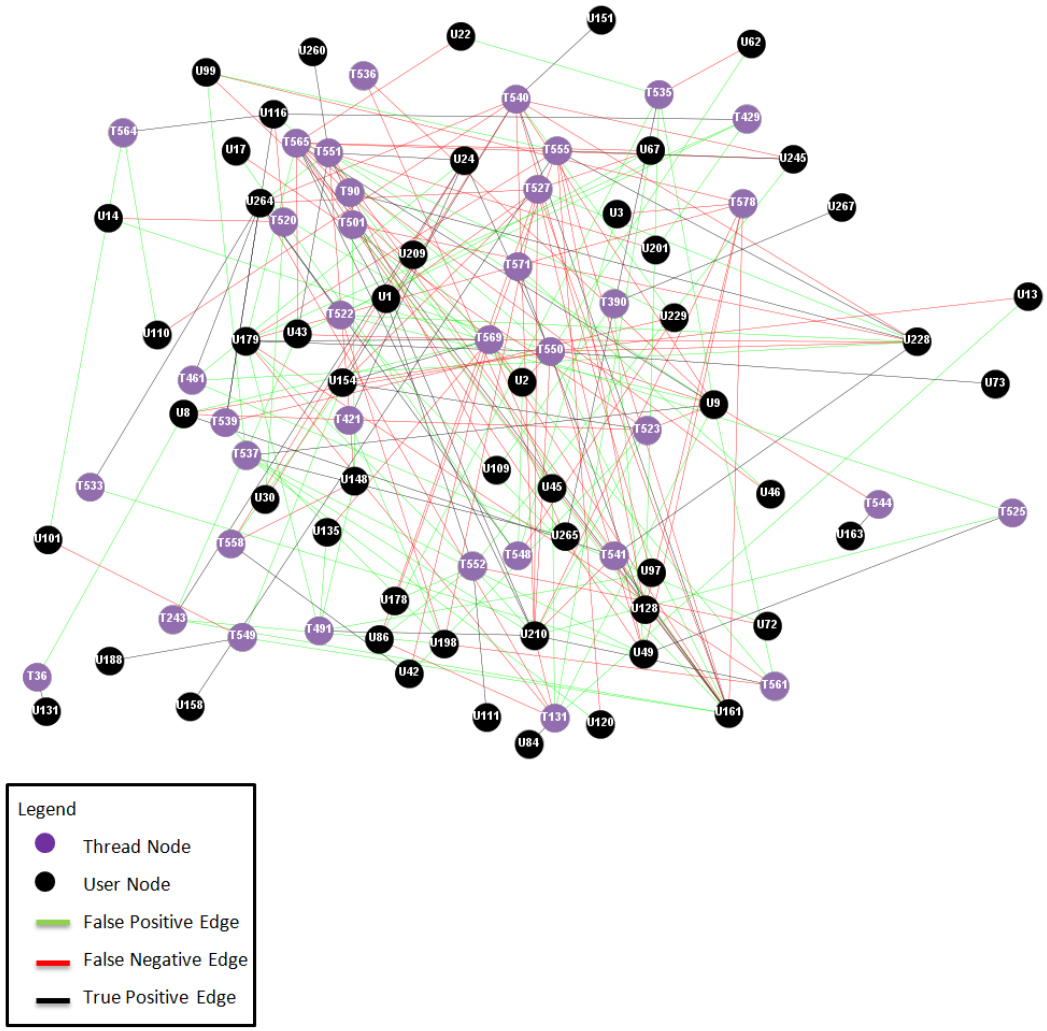


Figure 4.8: Network of Sub-Forum 5 for Month 6

Recall	Accuracy	Precision	F-measure
0.674	0.939	0.689	0.681

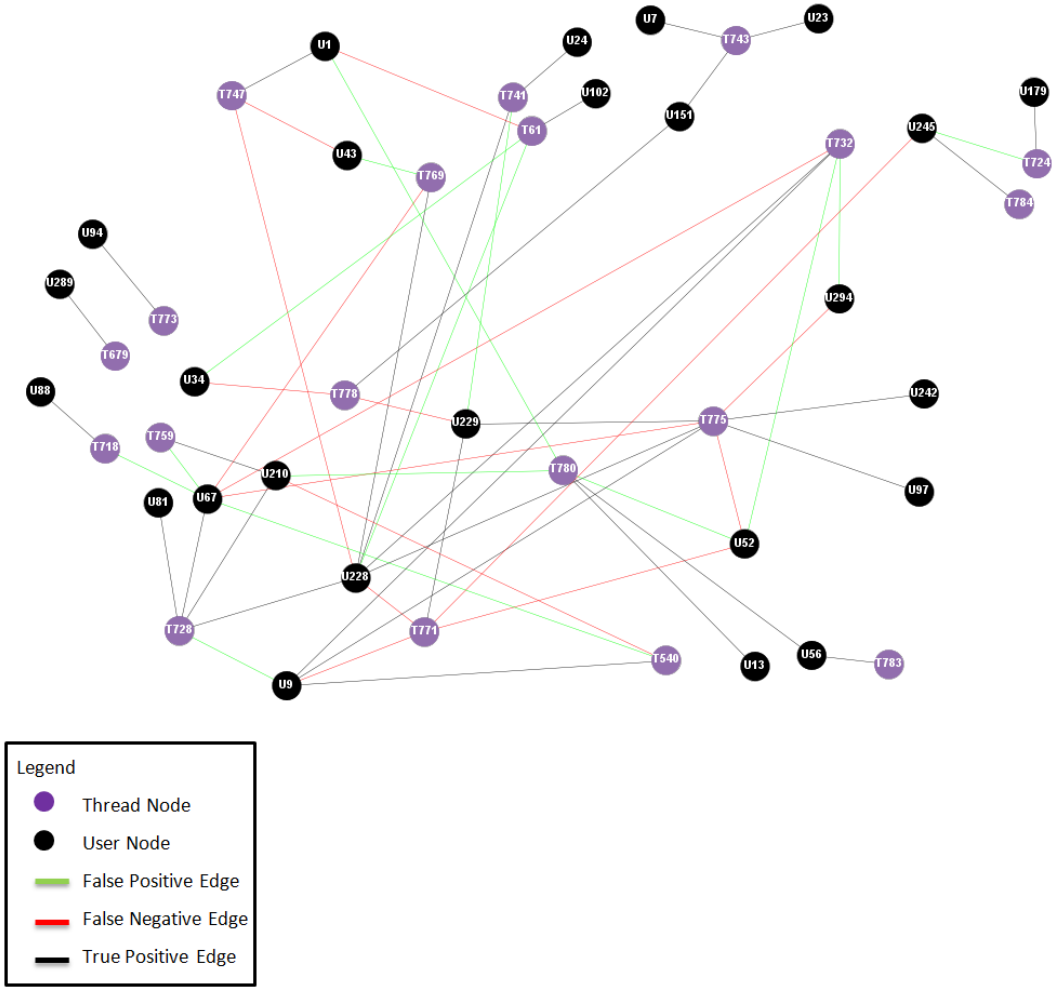


Figure 4.9: Network of Sub-Forum 5 for Month 9

4.3.5 Sub-Forum 6

In Table 4.11 we show the results obtained for each of the metrics evaluated for Sub-Forum 6. As we can notice, the best result with regards to F measure is obtained in month 10 and the worst in month 13. Note that we had problems with the data and could not run the model for month 6.

Table 4.11: Results of Sub-Forum 6

Month	**Recall**	**Accuracy**	**Precision**	**F-measure**
2	0.818	0.886	0.818	0.818
3	0.789	0.927	0.833	0.811
4	0.857	0.933	0.857	0.857
5	0.842	0.939	0.842	0.842
6*	-	-	-	-
7	0.842	0.914	0.842	0.842
8	0.800	0.852	0.800	0.800
9	0.895	0.961	0.944	0.919
10	0.947	0.973	0.947	0.947
11	0.846	0.900	0.846	0.846
12	0.842	0.933	0.842	0.842
13	0.647	0.848	0.733	0.688
Mean	**0.830**	**0.915**	**0.846**	**0.837**

Recall	Accuracy	Precision	F-measure
0.947	0.973	0.947	0.947

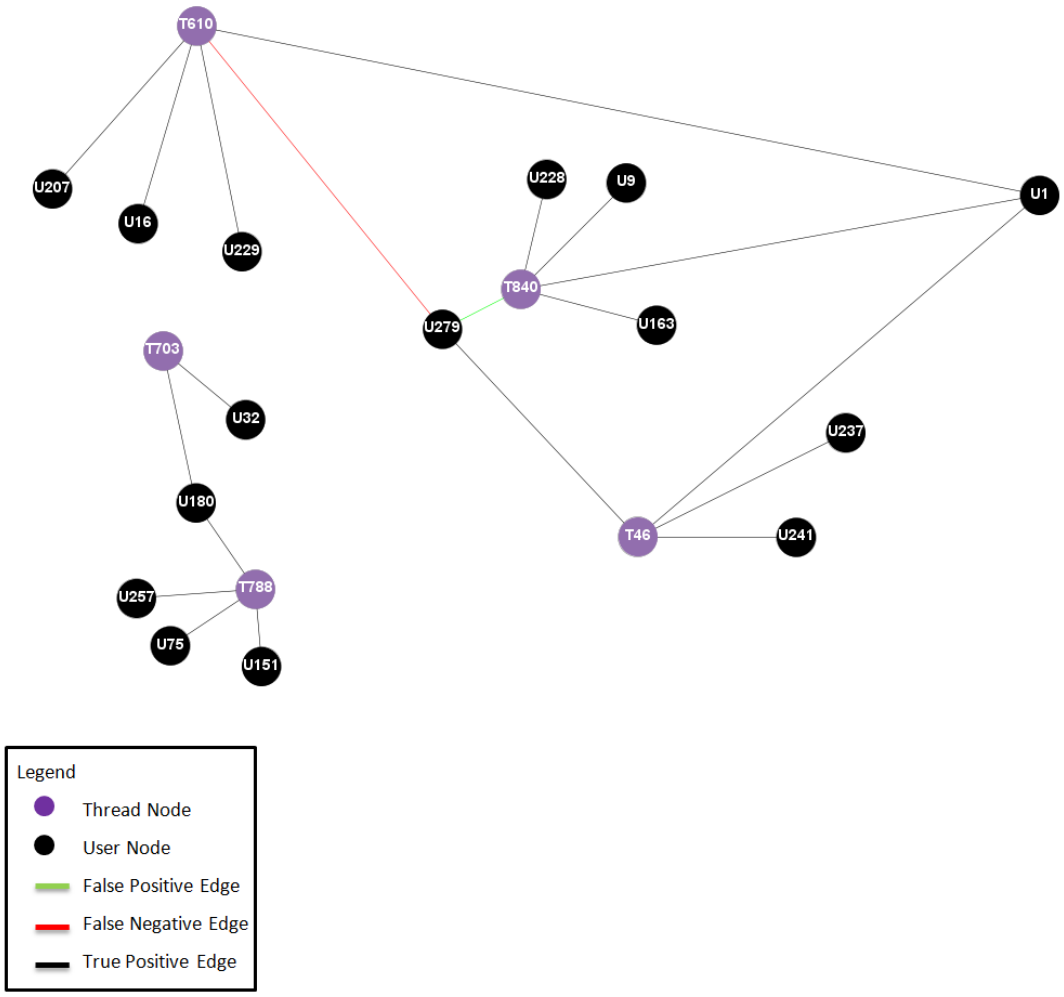


Figure 4.10: Network of Sub-Forum 6 for Month 10

Recall	Accuracy	Precision	F-measure
0.647	0.848	0.733	0.688

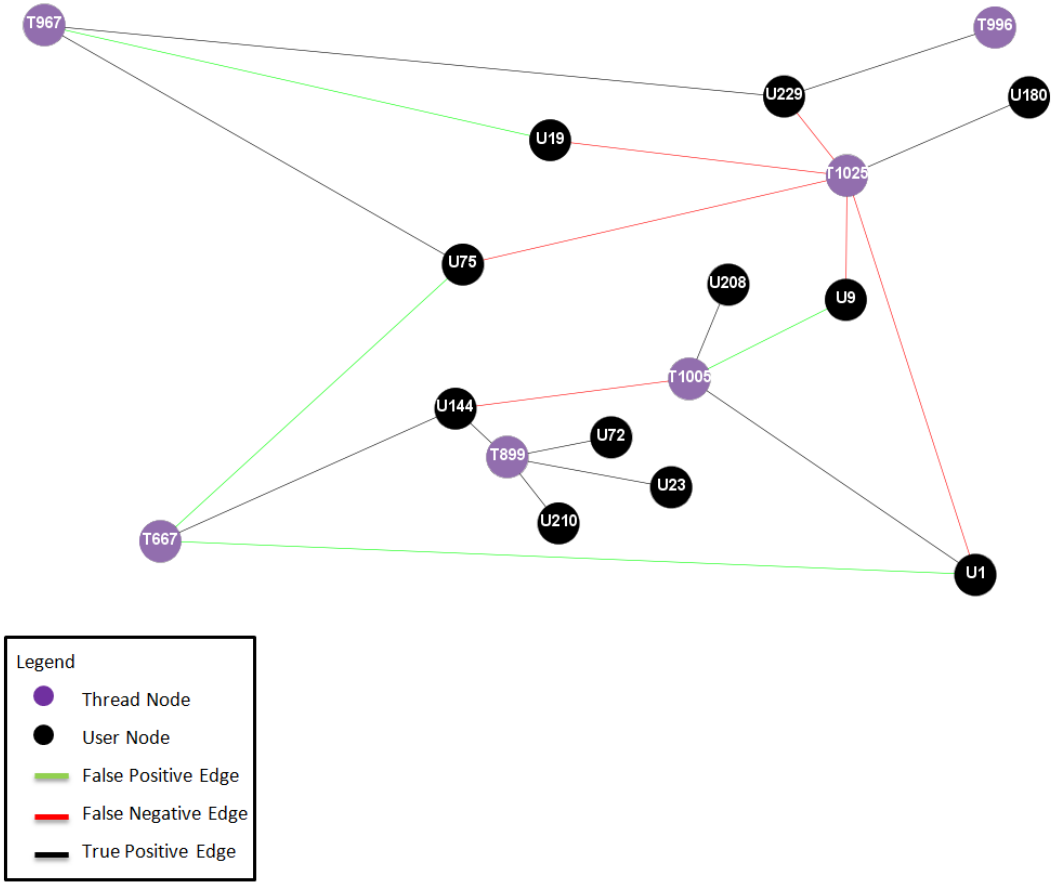


Figure 4.11: Network of Sub-Forum 6 for Month 13

4.4 Discussion

The proposed methodology with the LCA neurophysiological model achieves 61% of F measure score in average across all Sub-Forums successfully modeling the microscopic interactions of information diffusion with great accuracy. Therefore, the research hypothesis of this work has been validated. To the best of our knowl-

edge this has not been reported in literature before. We get the best results on Sub-Forum 6 where the low number of posts makes the content topics clearer and makes it easier for the model to find the threads a user finds interest in. We can observe that as the number of posts increases it becomes harder to distinguish where a user is motivated to post in based in content. In Fig. 4.12 we show the relationship between the number of posts made in a period of time (month) and the F measure score obtained by the model.

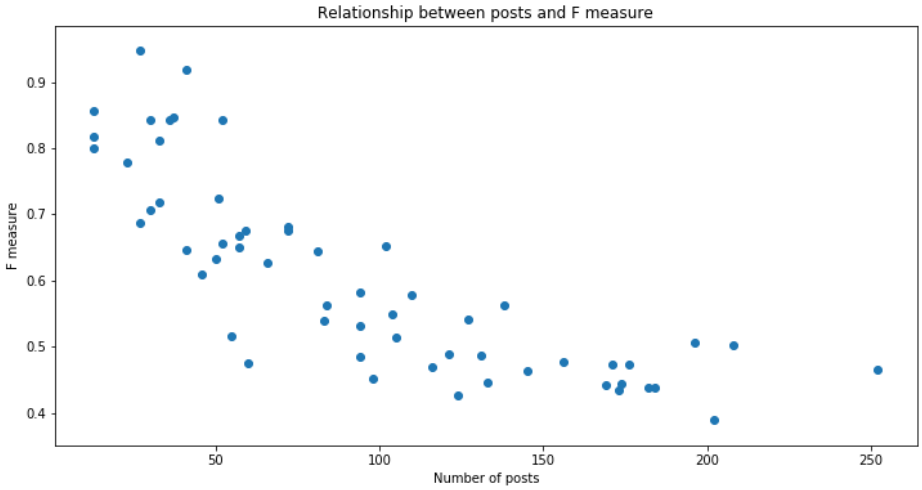


Figure 4.12: Relationship between number of posts and F-measure score

We hypothesize that this is caused by the level of noise (non-related content) that gets into the threads' text vector representations.

The number of posts in a period is directly related to the length chosen for the time period. Thus, choosing an optimal length for the time periods would be very beneficial for the results of the model. Another way in which we could enhance the model is to incorporate a discrimination behavior for users in which they do not consider posts which differ to much with the user's text preference vector, in a similar way to the one done in [46].

If we take a look to the temporal behavior of F measure within a Sub-Forum we can note that the scores do not deviate much from the mean value, this shows that our model is very robust in terms of temporal decay. This is probably associated to parameter a. In this research we set the value of $a = 50$ but this value is not

optimized. An interesting idea would be to incorporate this parameter into the optimization scheme. Another approach would be to test multiple functional forms to determine which captures user utility best.

The proposed network topology is also a success because, not only does it reduce the complexity of the problem as shown in section 3.2, but also manages to help network visualization by making less dense in terms of links. Besides, it helps capture the real nature in which a web user interacts with a web forum. Moreover, it allows for natural future extensions like different types of arcs, or node characteristics particular to certain type of nodes. This would allow us to include network topological features into the model and certainly help improve the results.

Overall, we accomplished all of our goals. We developed, implemented and tested a methodology that allows to deal with the information diffusion problem. We used and customized the LCA model for this purpose and obtained very successful results exceeding our expectations. We also contribute with a framework that can be used by future researchers interested in information diffusion.

Chapter 5

Conclusion

Today, web sites are evolving to social web sites. These sites are created to allow any user to share contents, work with each other, coordinate among many people, learn together, discuss a problem, etc. Besides, low entrance barriers allow any organization, enterprise or individual to have this type of web portals. Therefore, it is quite important to keep a well-structured and organized web site. However, to do so, we need the use of Social Network Analysis (SNA) and Web Mining (WM) techniques. in particular, understanding information diffusion within the web site can be of enormous help while sorting out the proper structure.

Although the field of information diffusion is a fairly new one, a lot of work has been done to attempt to model it. Most of this work has led to models that describe and predict the networks behavior in a successful manner but only at a macroscopic or mesoscopic level. Our work successfully managed to model diffusion at a microscopic level without losing accuracy. We propose to combine SNA with semantic based text mining techniques in order to fulfill two goals: first, to bridge the gap between SNA and WM approaches; and second, to attain better results when we apply any SNAs' metric or algorithms.

We proposed two fundamental changes in order to perform mentioned objectives: A network representation topology change and Topic-based text mining. In this research, Topic-based text mining was developed by using Latent Dirichlet Allocation (LDA). We applied our approach to predict link formation in the newly proposed topology in Plexilandia, which is a OSN with more than 15 years of life and 2500 members. Combining both changes allows a successful network reconstruction. Therefore, we successfully showed that the application of semantics into the SNA process is able to predict topological features of the network. Moreover, we showed that a content-driven approach is effective to model information

diffusion at a microscopic level.

As a summary, we developed a methodology for modeling information diffusion and stablished a framework available for future researchers. We accomplished all of the goals we established for this research, in particular, we exceeded our expectation obtaining 61% average F measure score. Finally, we validated our research hypothesis by showing that the LCA model is capable of modeling information diffusion at a microscopic level in a successful manner.

5.1 Future Work

As our first attempt to introduce semantics in the SNA process for the information diffusion problem through the use of WM techniques, although, it has been very successful, it seems necessary to make a deeper exploration into the basics of Natural Language Processing (NPL) algorithms. We believe that the NLP algorithms still lack in terms of capturing the real meaning of a text document because they focus on term frequency and joint occurrence of words in a document [66]. It would be possible to explore automatic ontology creation for a specific domain as a way to tackle this problem. When reviewing previous research, we can notice that the last survey about information diffusion was developed in 2013 so it would be beneficial for future researchers and for the area to generate an update of the work done in the field. Regarding the methodology proposed in this work there are some adjustments that could be implemented in the future such as

- Make a scheme that re-calibrates parameters in a stepwise fashion or every certain number of steps.
- Include a time period selection algorithm to be able to compute its optimal size. For example, we could apply a clustering algorithm over the threads' activation periods.
- Make a comparison between different functional forms to transform text vector representation similarity to user utility.
- Reformulate the habit formation modification to the LCA model in order to suit previous research, i.e. that habit formation follows an asymptotic curve for longer periods of time.
- Try other heuristic algorithms to make the calibration step less time consuming. Another possibility is to translate the code to another programming language.

- Implement and test other models of decision making.
- Extend the proposed network topology to include arcs between threads and between users in order to be able to capture more of the information available within the network.
- Combine the model with other existing models to, for example, predict the number of threads a user posts in, instead of using that as a model input.

Other areas where we need to advance, is in the creation of a distance between text vector representations that allows to extract the most value from the text content generated by users. Besides, we could try combining this work with existing methods to measure the improvement in prediction rate.

Bibliography

[1] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81.

[2] Gerard Salton, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing". In: *Communications of the ACM* 18.11 (1975), pp. 613–620.

[3] FM Bass. "A new product growth model for consumer durables, Mathematical Models in Marketing". In: *Lecture Notes in Economics and Mathematical Systems* 132 (1976), pp. 351–253.

[4] Mark Granovetter. "Threshold models of collective behavior". In: *American journal of sociology* 83.6 (1978), pp. 1420–1443.

[5] James L McClelland. "Toward a theory of information processing in graded, random, and interactive networks." In: (1993).

[6] Heinz Mühlenbein and Dirk Schlierkamp-Voosen. "Predictive models for the breeder genetic algorithm i. continuous parameter optimization". In: *Evolutionary computation* 1.1 (1993), pp. 25–49.

[7] Colin R Reeves. "Genetic algorithms and neighbourhood search". In: *AISB Workshop on Evolutionary Computing*. Springer. 1994, pp. 115–130.

[8] Mandavilli Srinivas and Lalit M Patnaik. "Genetic algorithms: A survey". In: *computer* 27.6 (1994), pp. 17–26.

[9] Barry Wellman et al. "Computer networks as social networks: Collaborative work, telework, and virtual community". In: *Annual review of sociology* 22.1 (1996), pp. 213–238.

[10] Barry Wellman and Milena Gulia. "Virtual communities as communities". In: *Communities in cyberspace* (1999), pp. 167–194.

[11] Amy Jo Kim. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[12] Lada A Adamic et al. "Search in power-law networks". In: *Physical review E* 64.4 (2001), p. 046135.

[13] Jacob Goldenberg, Barak Libai, and Eitan Muller. "Talk of the network: A complex systems look at the underlying process of word-of-mouth". In: *Marketing letters* 12.3 (2001), pp. 211–223.

[14] Christopher M Johnson. "A survey of current research on online communities of practice". In: *The internet and higher education* 4.1 (2001), pp. 45–60.

[15] Marius Usher and James L McClelland. "The time course of perceptual choice: the leaky, competing accumulator model." In: *Psychological review* 108.3 (2001), p. 550.

[16] Barry Wellman. "Computer networks as social networks". In: *Science* 293.5537 (2001), pp. 2031–2034.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[18] David Kempe, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 137–146.

[19] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[20] Aron Culotta, Ron Bekkerman, and Andrew McCallum. *Extracting social networks and contact information from email and the web*. Tech. rep. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[21] Eyal Even-Dar and Asaf Shapira. "A note on maximizing the spread of influence in social networks". In: *International Workshop on Web and Internet Economics*. Springer. 2007, pp. 281–286.

[22] Joshua I Gold and Michael N Shadlen. "The neural basis of decision making". In: *Annual review of neuroscience* 30 (2007).

[23] Masao Kubo et al. "The possibility of an epidemic meme analogy for web community population analysis". In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2007, pp. 1073–1080.

[24] David Mimno, Hanna Wallach, and Andrew McCallum. "Community-based link prediction with text". In: *Proc. of NIPS*. 2007.

[25] Sebastián A Rıos. "A study on web mining techniques for off-line enhancements of web sites". PhD thesis. 2007.

[26] Xiaodan Song et al. "Information flow modeling based on diffusion rate for prediction and ranking". In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 191–200.

[27] Dongshan Xing and Mark Girolami. "Employing Latent Dirichlet Allocation for fraud detection in telecommunications". In: *Pattern Recognition Letters* 28.13 (2007), pp. 1727–1734.

[28] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking". In: *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE. 2008, pp. 3–12.

[29] X. H. Phang and CT. Nguyen. "Gibbslda++". In: (2008).

[30] Haibo Hu and Xiaofan Wang. "Evolution of a large online social network". In: *Physics Letters A* 373.12-13 (2009), pp. 1105–1110.

[31] Sebastián A Ríos, Felipe Aguilera, and Luis A Guerrero. "Virtual communities of practice's purpose evolution analysis using a concept-based mining approach". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2009, pp. 480–489.

[32] Noga Alon et al. "A note on competitive diffusion through social networks". In: *Information Processing Letters* 110.6 (2010), pp. 221–225.

[33] Héctor Alvarez. "Detección de miembros clave en una comunidad virtual de prática mediante análisis de redes sociales y mineria de datos avanzada". In: *Master's thesis, University of Chile* (2010).

[34] Héctor Alvarez et al. "Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2010, pp. 591–600.

[35] GASTON ANDRÉS L'HUILLIER CHAPARRO et al. "CLASIFICACION DE PHISHING UTILIZANDO MINERÍA DE DATOS ADVERSARIAL Y JUEGOS CON INFORMACION INCOMPLETA". In: (2010).

[36] Maksim Kitsak et al. "Identification of influential spreaders in complex networks". In: *Nature physics* 6.11 (2010), p. 888.

[37] Phillippa Lally et al. "How are habits formed: Modelling habit formation in the real world". In: *European journal of social psychology* 40.6 (2010), pp. 998–1009.

[38] Eduardo Merlo et al. "Finding inner copy communities using social network analysis". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2010, pp. 581–590.

[39] Mohammad Al Hasan and Mohammed J Zaki. "A survey of link prediction in social networks". In: *Social network data analytics*. Springer, 2011, pp. 243–275.

[40] Phil E Brown and Junlan Feng. "Measuring user influence on twitter using modified k-shell decomposition". In: *Fifth international AAAI conference on weblogs and social media*. 2011.

[41] Lautaro Cuadra, Sebastín A Rios, and Gaston L'Huillier. "Enhancing community discovery and characterization in vcop using topic models". In: *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*. IEEE Computer Society. 2011, pp. 326–329.

[42] Conrad Lee, Thomas Scherngell, and Michael J Barber. "Investigating an online social network using spatial interaction models". In: *Social Networks* 33.2 (2011), pp. 129–133.

[43] Gastón L'huillier et al. "Topic-based social network analysis for virtual communities of interests in the dark web". In: *ACM SIGKDD Explorations Newsletter* 12.2 (2011), pp. 66–73.

[44] Jiyoung Woo, Jaebong Son, and Hsinchun Chen. "An SIR model for violent topic diffusion in social media". In: *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*. IEEE. 2011, pp. 15–19.

[45] Rakesh Kumar. "Blending roulette wheel selection & rank selection in genetic algorithms". In: *International Journal of Machine Learning and Computing* 2.4 (2012), p. 365.

[46] Lin Li et al. "Phase transition in opinion diffusion in social networks". In: *Acoustics, speech and signal processing (ICASSP), 2012 IEEE international conference on*. IEEE. 2012, pp. 3073–3076.

[47] Seth A Myers, Chenguang Zhu, and Jure Leskovec. "Information diffusion and external influence in networks". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 33–41.

[48] Pablo E Román, Miguel E Gutiérrez, and Sebastián A Rios. "A model for content generation in On-line social network." In: *KES*. 2012, pp. 756–765.

[49] Reiko Takehara, Masahiro Hachimori, and Maiko Shigeno. "A comment on pure-strategy Nash equilibria in competitive diffusion games". In: *Information processing letters* 112.3 (2012), pp. 59–60.

[50] Jiyoung Woo and Hsinchun Chen. "An event-driven SIR model for topic diffusion in web forums". In: *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE. 2012, pp. 108–113.

[51] Fei Xiong et al. "An information diffusion model based on retweeting mechanism for online social media". In: *Physics Letters A* 376.30-31 (2012), pp. 2103–2108.

[52] Adrien Guille et al. "Information diffusion in online social networks: A survey". In: *ACM Sigmod Record* 42.2 (2013), pp. 17–28.

[53] Adrien Guille et al. "Sondy: An open source platform for social dynamics mining and analysis". In: *Proceedings of the 2013 ACM SIGMOD international conference on management of data*. ACM. 2013, pp. 1005–1008.

[54] Jianwei Niu et al. "An Empirical Study of a Chinese Online Social Network–Renren". In: *Computer* 46.9 (2013), pp. 78–84.

[55] Lucy Small and Oliver Mason. "Information diffusion on the iterated local transitivity model of online social networks". In: *Discrete Applied Mathematics* 161.10-11 (2013), pp. 1338–1344.

[56] Lucy Small and Oliver Mason. "Nash equilibria for competitive information diffusion on trees". In: *Information Processing Letters* 113.7 (2013), pp. 217–219.

[57] Chunxiao Jiang, Yan Chen, and KJ Ray Liu. "Modeling information diffusion dynamics over social networks". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 1095–1099.

[58] Sebastián A Ríos and Ricardo Muñoz. "Content patterns in topic-based overlapping communities". In: *The Scientific World Journal* 2014 (2014).

[59] Ye Sun et al. "Epidemic spreading on weighted complex networks". In: *Physics Letters A* 378.7-8 (2014), pp. 635–640.

[60] Li-Jen Kao and Yo-Ping Huang. "Mining influential users in social network". In: *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE. 2015, pp. 1209–1214.

[61] Chuan Luo, Xiaolong Zheng, and Daniel Zeng. "Inferring social influence and meme interaction with Hawkes processes". In: *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*. IEEE. 2015, pp. 135–137.

[62] Akrati Saxena, SRS Iyengar, and Yayati Gupta. "Understanding spreading patterns on social networks based on network topology". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. IEEE. 2015, pp. 1616–1617.

[63] Anupriya Shukla, Hari Mohan Pandey, and Deepti Mehrotra. "Comparative review of selection techniques in genetic algorithm". In: *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on*. IEEE. 2015, pp. 515–519.

[64] John Breslin Tope Omitola Ríos Sebastián. *Social Semantic Web Intelligence*. Morgan & Claypool Publishers, 2015.

[65] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. "Evaluating link prediction methods". In: *Knowledge and Information Systems* 45.3 (2015), pp. 751–782.

[66] Constanza Contreras-Piña and Sebastián A Ríos. "An empirical comparison of latent sematic models for applications in industry". In: *Neurocomputing* 179 (2016), pp. 176–185.

[67] Dong Li et al. "Exploiting information diffusion feature for link prediction in sina weibo". In: *Scientific reports* 6 (2016), p. 20058.

[68] Xiaoyan Qiu et al. "Effects of time-dependent diffusion behaviors on the rumor spreading in social networks". In: *Physics Letters A* 380.24 (2016), pp. 2054–2063.

[69] Jiyoung Woo and Hsinchun Chen. "Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog". In: *SpringerPlus* 5.1 (2016), p. 66.

[70] Ying Hu, Rachel Jeungeun Song, and Min Chen. "Modeling for information diffusion in online social networks via hydrodynamics". In: *IEEE Access* 5 (2017), pp. 128–135.

[71] Sebastián A Ríos et al. "Semantically enhanced network analysis for influencer identification in online social networks". In: *Neurocomputing* (2017).

[72] Hadi Shakibian and Nasrollah Moghadam Charkari. "Mutual information model for link prediction in heterogeneous complex networks". In: *Scientific Reports* 7 (2017), p. 44981.

[73] Jiawei Zhang et al. "Link prediction with cardinality constraint". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM. 2017, pp. 121–130.

[74] Ricardo Baeza-Yates. "Bias on the web". In: *Communications of the ACM* 61.6 (2018), pp. 54–61.

Annex A

Network Images

A.1 Remaining Sub-Forum 2 Network Images

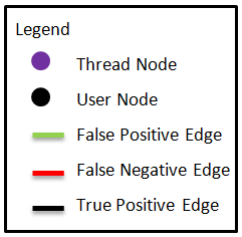
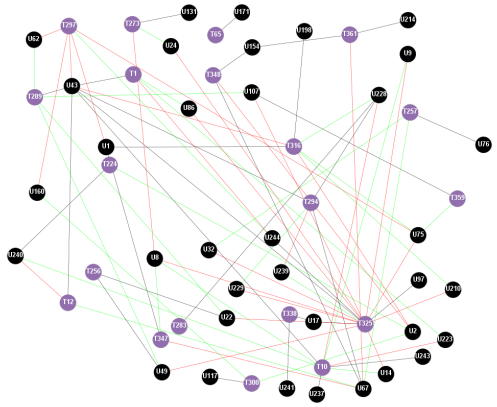
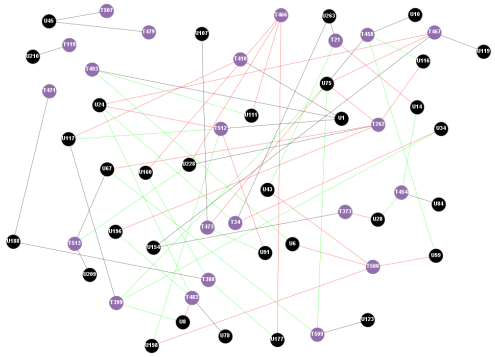


Figure A.1: Legend for network graphs

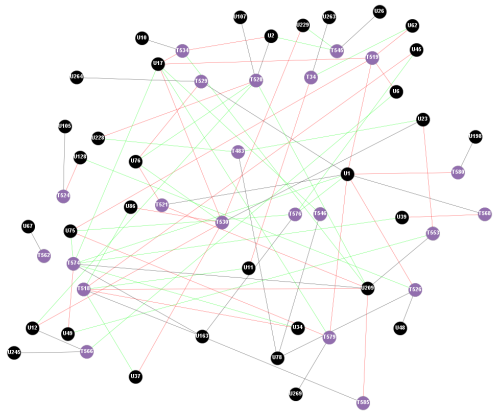


(a) Sub-Forum 2 Month 3

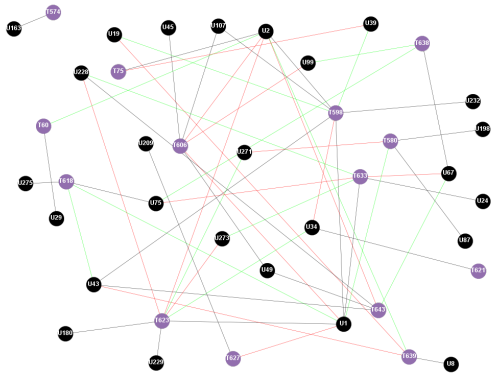


(b) Sub-Forum 2 Month 5

Figure A.2: Networks of Sub-Forum 2 for (a) Month 3 and (b) Month 5

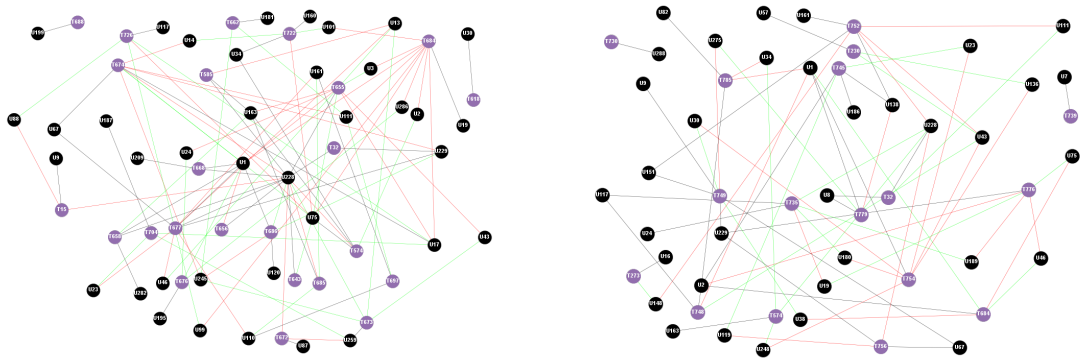


(a) Sub-Forum 2 Month 6



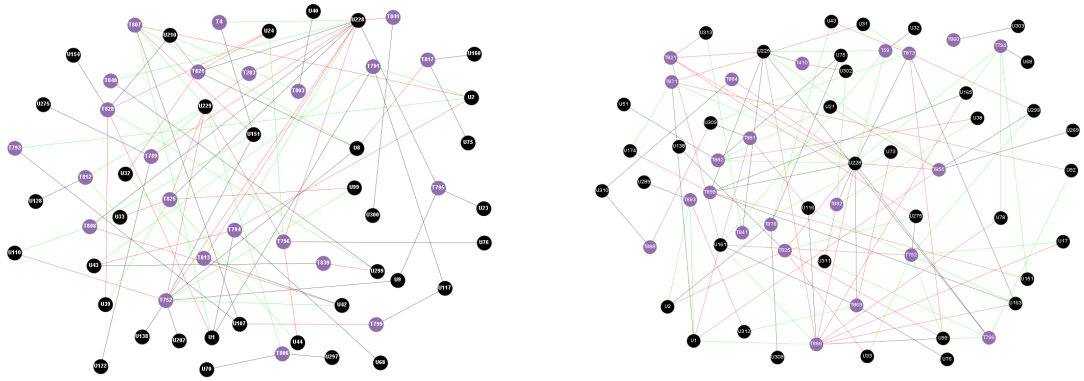
(b) Sub-Forum 2 Month 7

Figure A.3: Networks of Sub-Forum 2 for (a) Month 6 and (b) Month 7



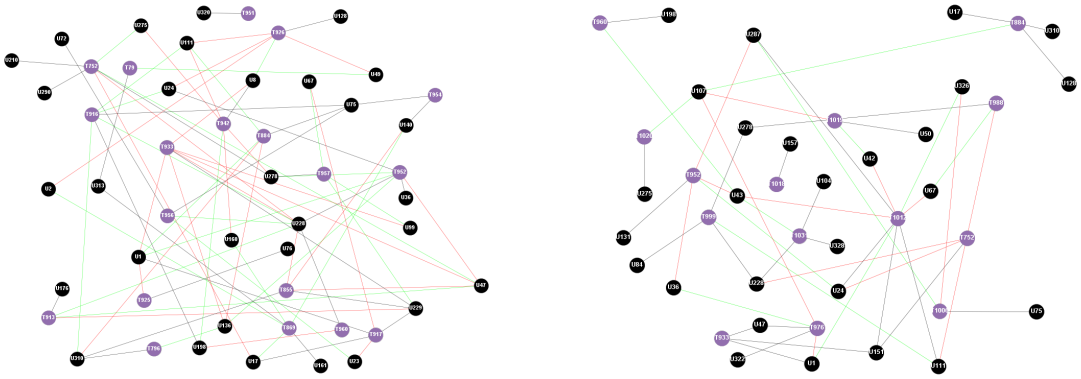
(a) Sub-Forum 2 Month 8 (b) Sub-Forum 2 Month 9

Figure A.4: Networks of Sub-Forum 2 for (a) Month 8 and (b) Month 9



(a) Sub-Forum 2 Month 10 (b) Sub-Forum 2 Month 11

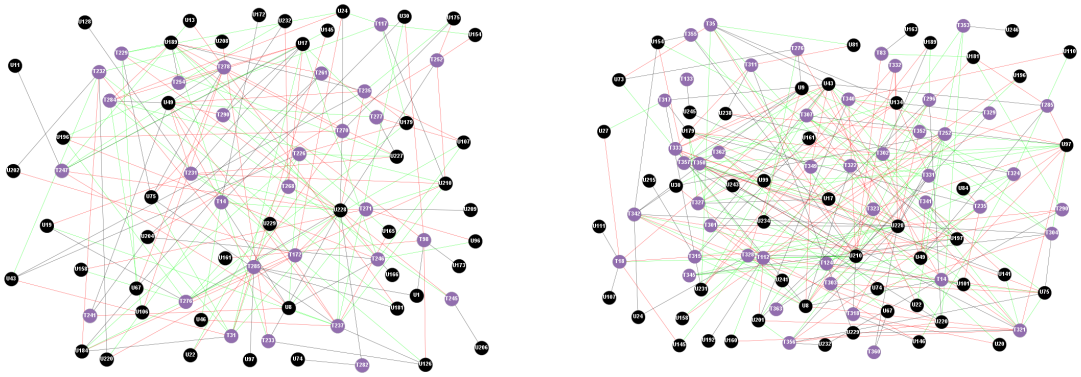
Figure A.5: Networks of Sub-Forum 2 for (a) Month 10 and (b) Month 11



(a) Sub-Forum 2 Month 12 (b) Sub-Forum 2 Month 13

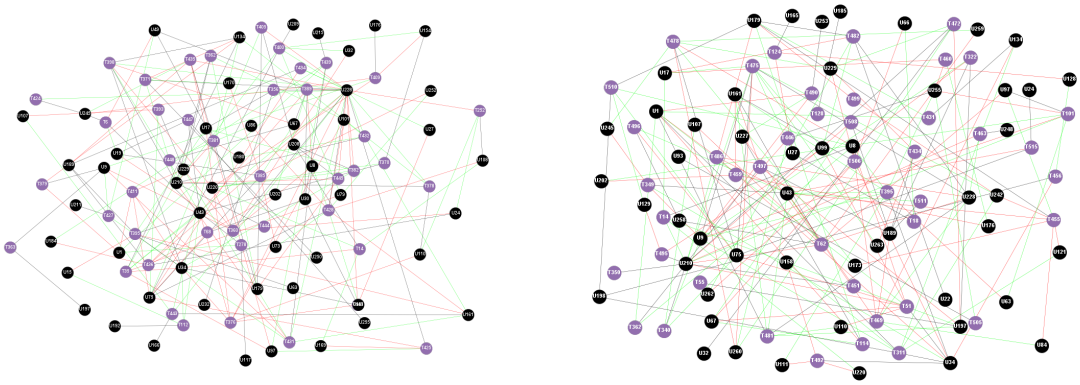
Figure A.6: Networks of Sub-Forum 2 for (a) Month 12 and (b) Month 13

A.2 Remaining Sub-Forum 3 Network Images



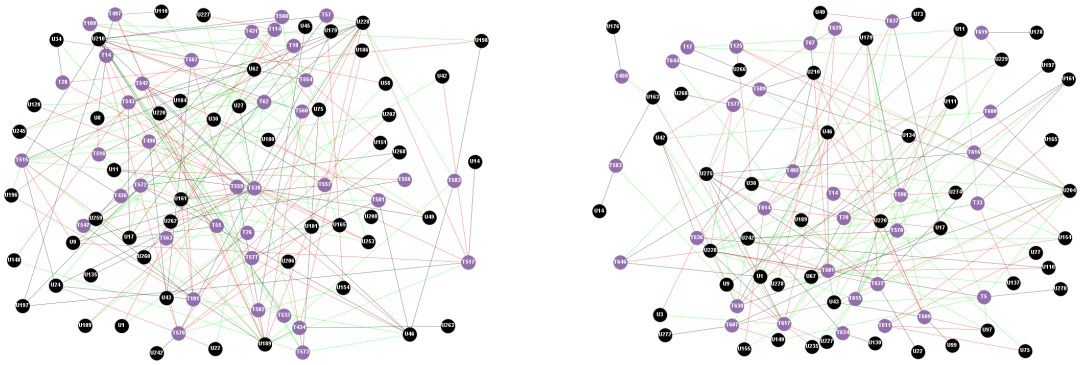
(a) Sub-Forum 3 Month 2 (b) Sub-Forum 3 Month 3

Figure A.7: Networks of Sub-Forum 3 for (a) Month 2 and (b) Month 3



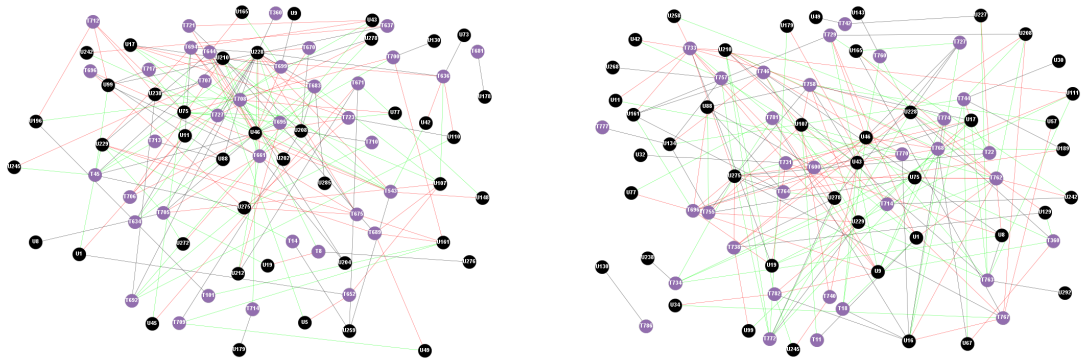
(a) Sub-Forum 3 Month 4 (b) Sub-Forum 3 Month 5

Figure A.8: Networks of Sub-Forum 3 for (a) Month 4 and (b) Month 5



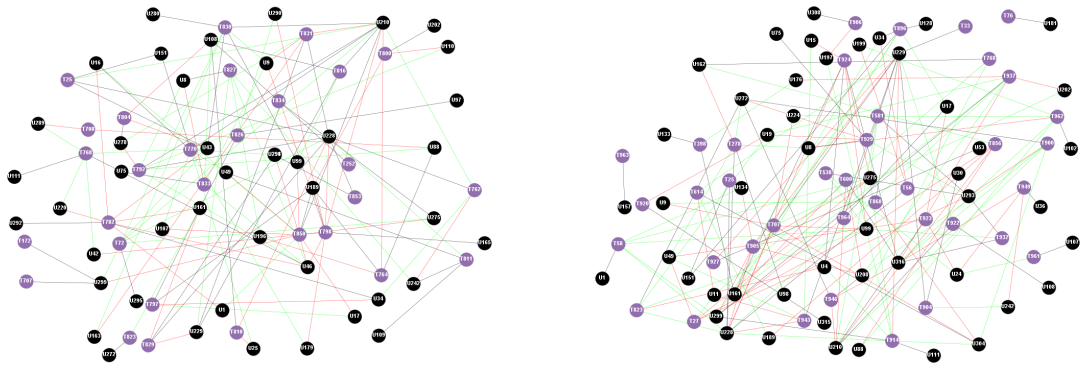
(a) Sub-Forum 3 Month 6 (b) Sub-Forum 3 Month 7

Figure A.9: Networks of Sub-Forum 3 for (a) Month 6 and (b) Month 7



(a) Sub-Forum 3 Month 8 (b) Sub-Forum 3 Month 9

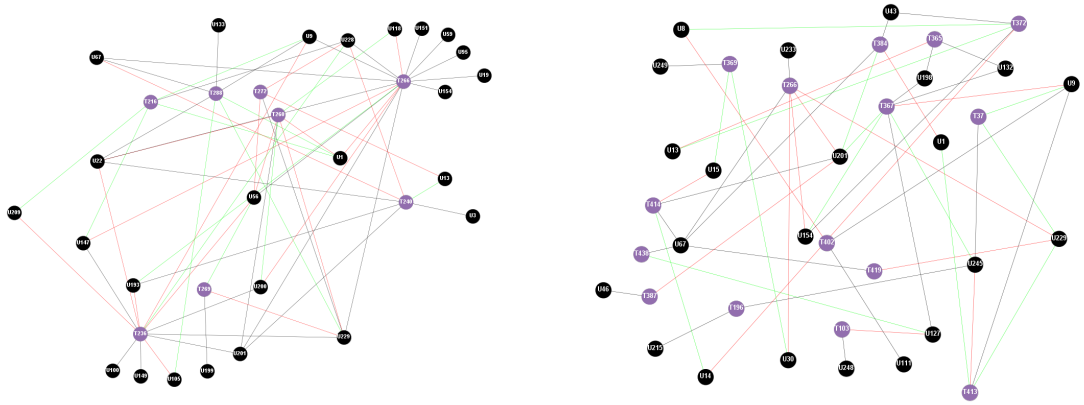
Figure A.10: Networks of Sub-Forum 3 for (a) Month 8 and (b) Month 9



(a) Sub-Forum 3 Month 10 (b) Sub-Forum 3 Month 12

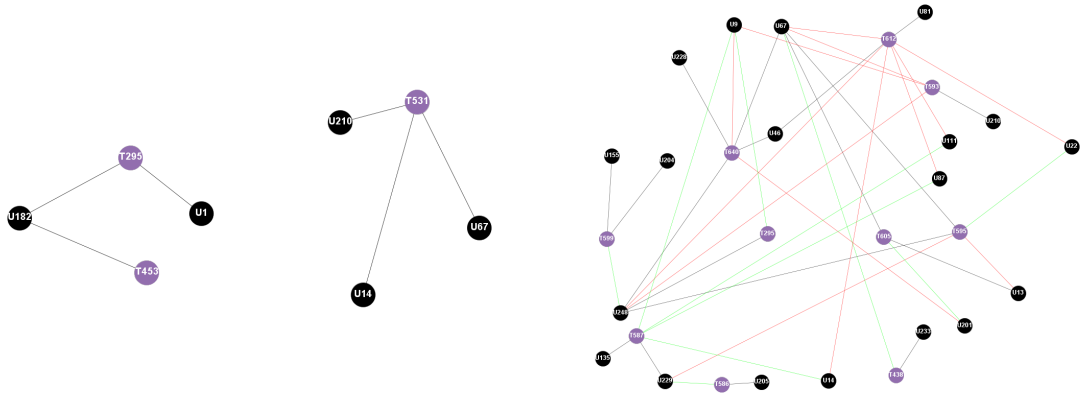
Figure A.11: Networks of Sub-Forum 3 for (a) Month 10 and (b) Month 12

A.3 Remaining Sub-Forum 4 Network Images



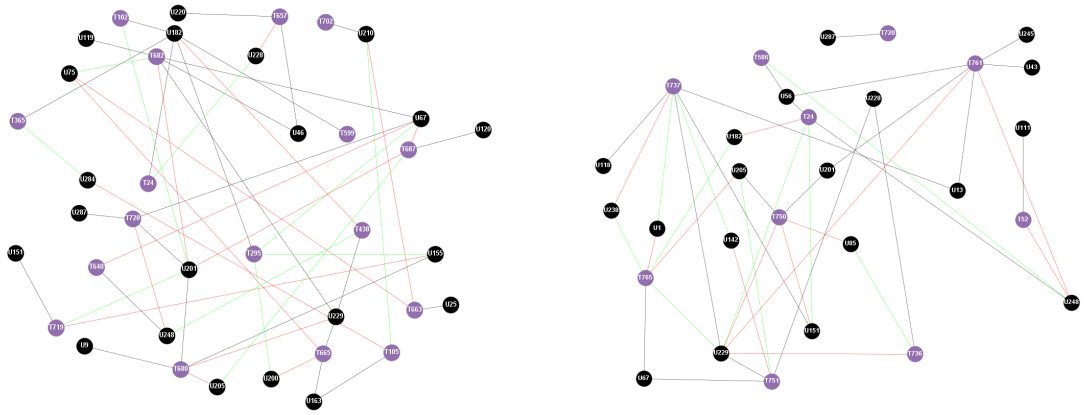
(a) Sub-Forum 4 Month 2 (b) Sub-Forum 4 Month 4

Figure A.12: Networks of Sub-Forum 4 for (a) Month 2 and (b) Month 4



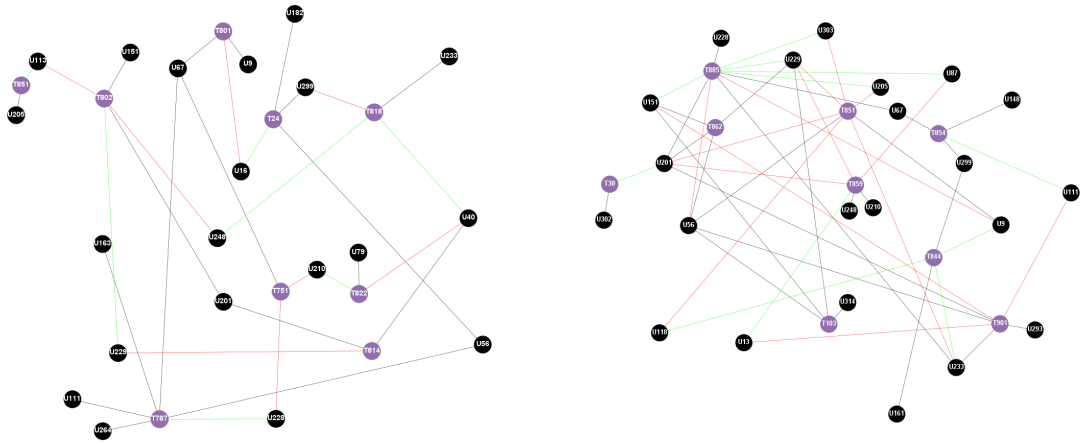
(a) Sub-Forum 4 Month 6 (b) Sub-Forum 4 Month 7

Figure A.13: Networks of Sub-Forum 4 for (a) Month 6 and (b) Month 7



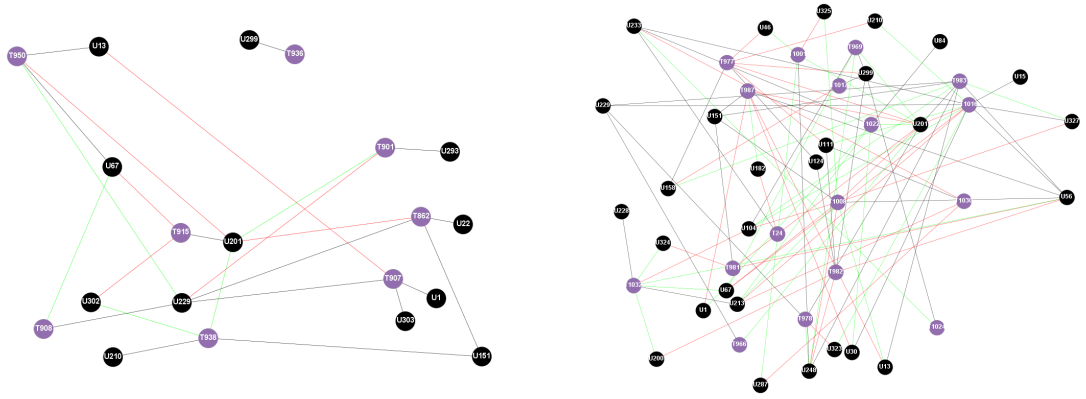
(a) Sub-Forum 4 Month 8 (b) Sub-Forum 4 Month 9

Figure A.14: Networks of Sub-Forum 4 for (a) Month 8 and (b) Month 9



(a) Sub-Forum 4 Month 10 (b) Sub-Forum 4 Month 11

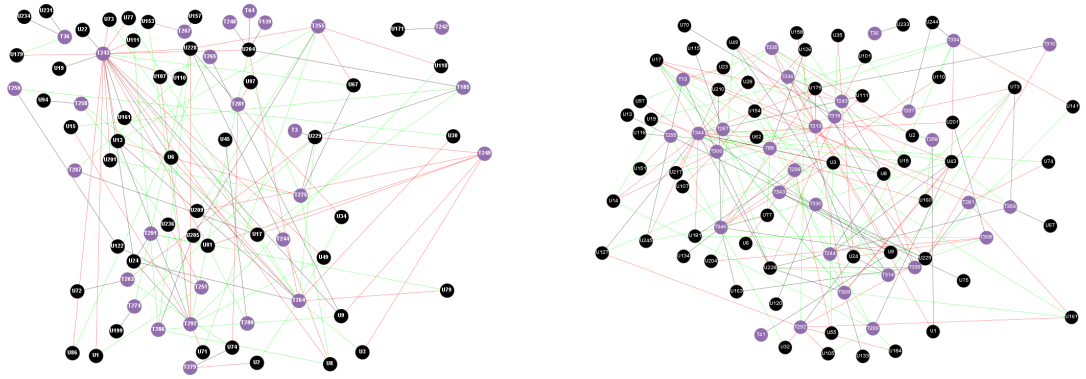
Figure A.15: Networks of Sub-Forum 4 for (a) Month 10 and (b) Month 11



(a) Sub-Forum 4 Month 12 (b) Sub-Forum 4 Month 13

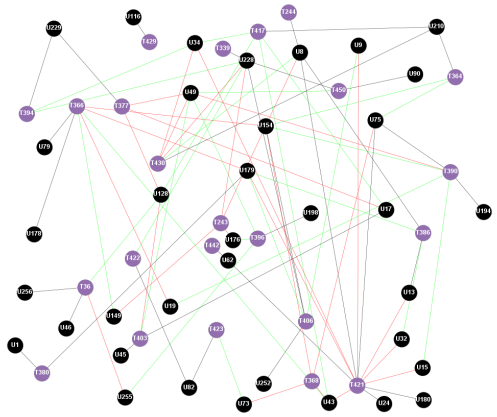
Figure A.16: Networks of Sub-Forum 4 for (a) Month 12 and (b) Month 13

A.4 Remaining Sub-Forum 5 Network Images

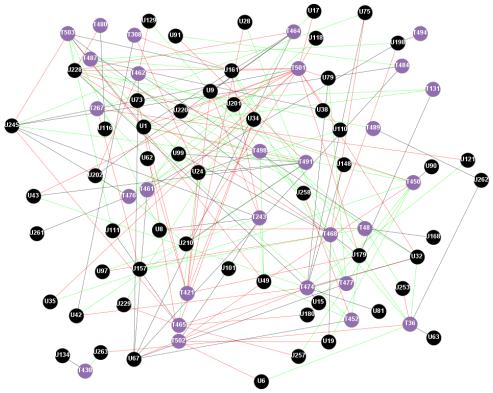


(a) Sub-Forum 5 Month 2 (b) Sub-Forum 5 Month 3

Figure A.17: Networks of Sub-Forum 5 for (a) Month 2 and (b) Month 3

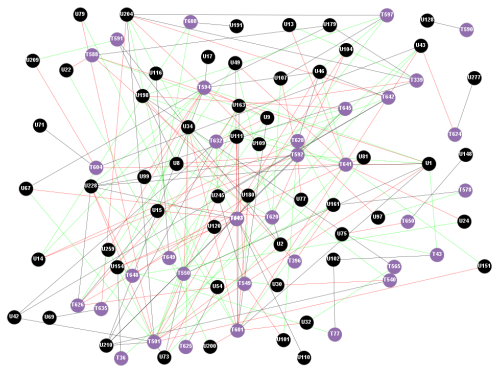


(a) Sub-Forum 5 Month 4

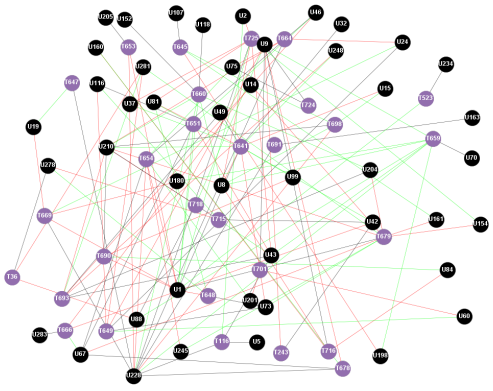


(b) Sub-Forum 5 Month 5

Figure A.18: Networks of Sub-Forum 5 for (a) Month 4 and (b) Month 5

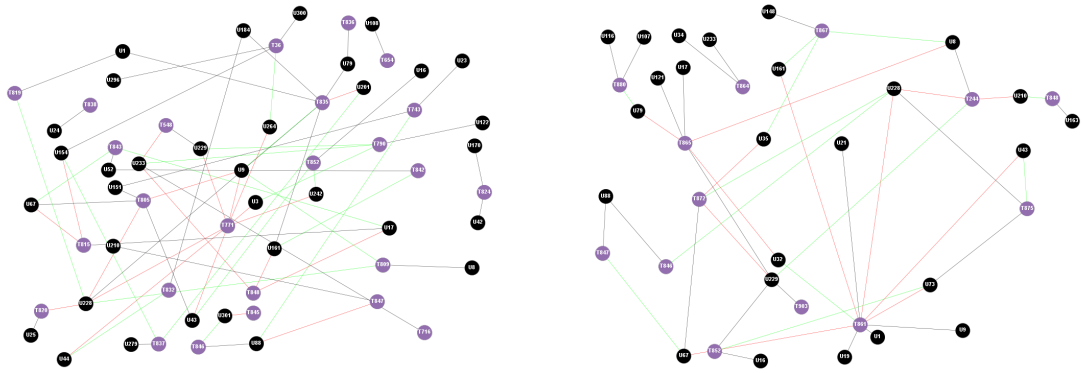


(a) Sub-Forum 5 Month 7



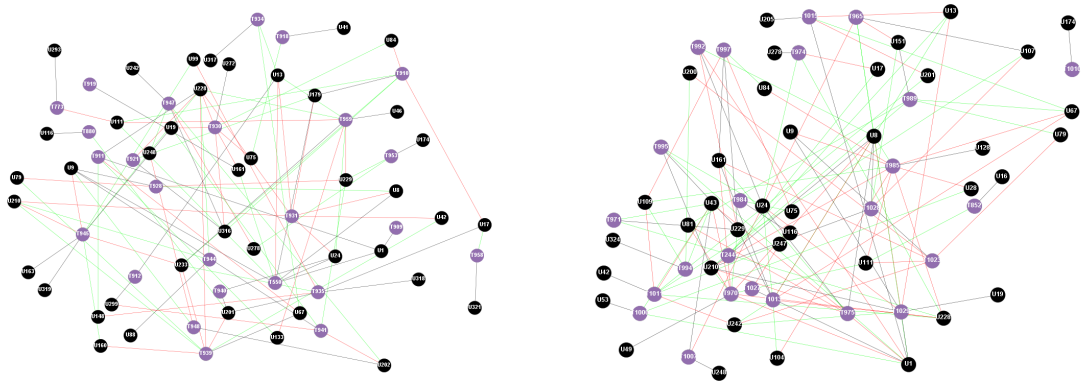
(b) Sub-Forum 5 Month 8

Figure A.19: Networks of Sub-Forum 5 for (a) Month 7 and (b) Month 8



(a) Sub-Forum 5 Month 10 (b) Sub-Forum 5 Month 11

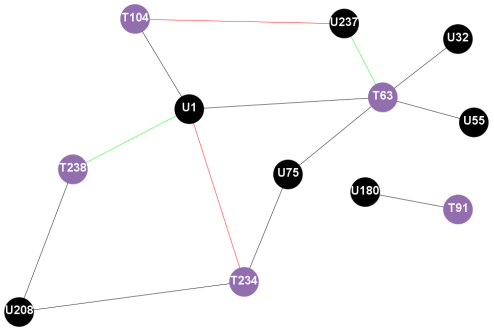
Figure A.20: Networks of Sub-Forum 5 for (a) Month 10 and (b) Month 11



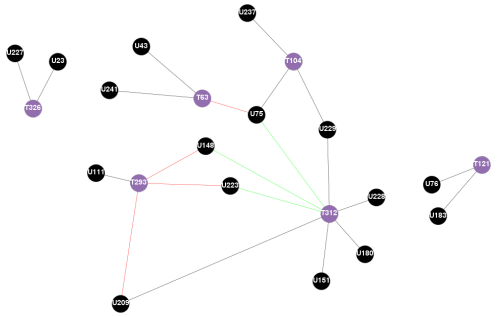
(a) Sub-Forum 5 Month 12 (b) Sub-Forum 5 Month 13

Figure A.21: Networks of Sub-Forum 5 for (a) Month 12 and (b) Month 13

A.5 Remaining Sub-Forum 6 Network Images

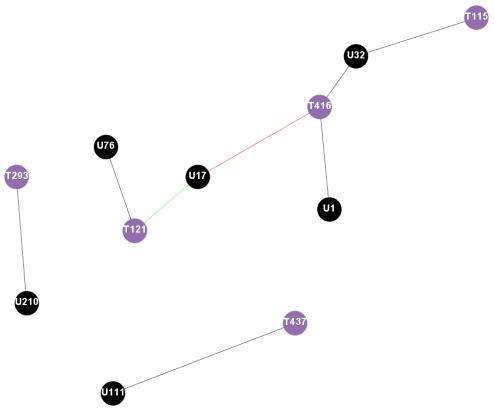


(a) Sub-Forum 6 Month 2

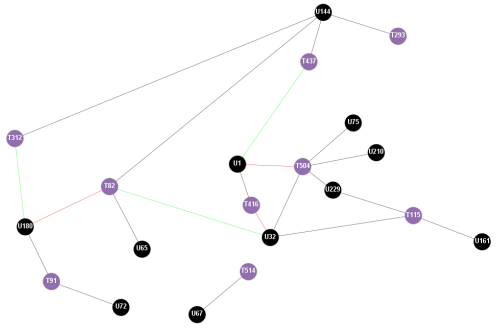


(b) Sub-Forum 6 Month 3

Figure A.22: Networks of Sub-Forum 6 for (a) Month 2 and (b) Month 3

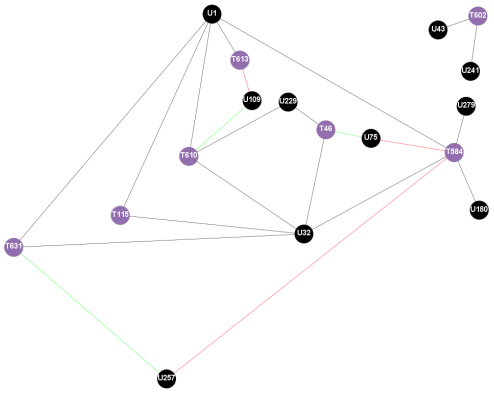


(a) Sub-Forum 6 Month 4

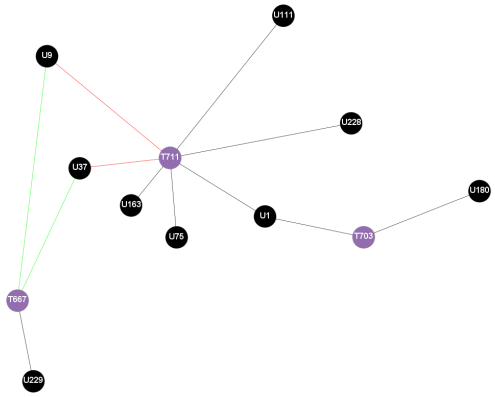


(b) Sub-Forum 6 Month 5

Figure A.23: Networks of Sub-Forum 6 for (a) Month 4 and (b) Month 5

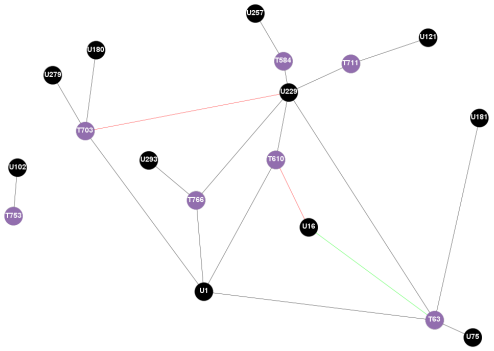


(a) Sub-Forum 6 Month 7

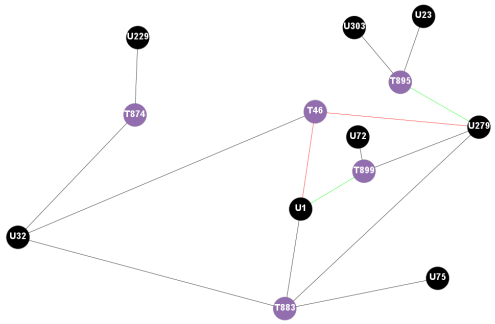


(b) Sub-Forum 6 Month 8

Figure A.24: Networks of Sub-Forum 6 for (a) Month 7 and (b) Month 8



(a) Sub-Forum 6 Month 9



(b) Sub-Forum 6 Month 11

Figure A.25: Networks of Sub-Forum 6 for (a) Month 9 and (b) Month 11

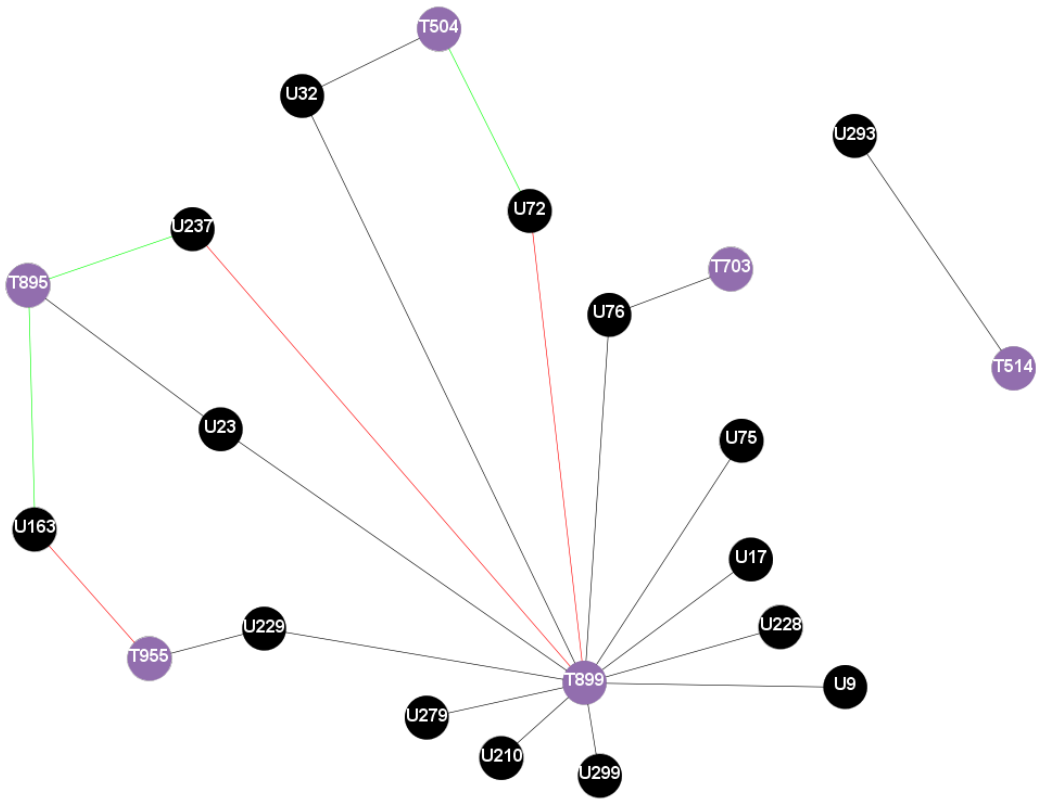


Figure A.26: Network of Sub-Forum 6 for Month 12

A.6 Others

Table A.1: β calibrated values

Sub-Forum	$\beta_{\mathbf{A}}$	$\beta_{\mathbf{B}}$	$\beta_{\mathbf{C}}$	$\beta_{\mathbf{X}}$
2	0.8627908535	0.147898369096	0.51057168026	0.55277821110
3	0.58428296434	0.90616373636	0.388960210471	0.0293256700512
4	0.58581658629	0.83305375118	0.35208228175	0.476104983946
5	0.62756874613	0.183592791154	0.0000227721	0.42883467851
6	0.5155427342	0.126055356488	0.48970952965	0.59505459800

Table A.2: κ calibrated values

Sub-Forum	$\kappa_{\mathbf{A}}$	$\kappa_{\mathbf{B}}$	$\kappa_{\mathbf{C}}$	$\kappa_{\mathbf{X}}$
2	0.174364160859	0.054631273654	0.069565857305	0.96545895734
3	0.68422639187	0.34001726318	0.21705810345	0.58816127791
4	0.64234099417	0.389164708224	0.86625978524	0.980585167117
5	0.70704911061	0.73300029795	0.046703670426	0.62252476318
6	0.28723792245	0.69171441868	0.087239533785	0.40065358219

Table A.3: λ calibrated values

Sub-Forum	$\lambda_{\mathbf{A}}$	$\lambda_{\mathbf{B}}$	$\lambda_{\mathbf{C}}$	$\lambda_{\mathbf{X}}$
2	0.491372623424	0.136918214609	0.39937333845	0.18918673259
3	0.145639007382	0.95054858333	0.189498650367	0.94880733805
4	0.63908203106	0.47785001044	0.107069371413	0.244599631189
5	0.093499394537	0.86386901172	0.84696142577	0.63967912728
6	0.956165857	0.86898713482	0.0439818761517	0.31492402604

Table A.4: F-measure Results

Month	**SF 2**	**SF 3**	**SF 4**	**SF 5**	**SF 6**
2	0.724	0.442	0.645	0.487	0.818
3	0.539	0.465	0.476	0.439	0.811
4	0.446	0.506	0.632	0.562	0.857
5	0.517	0.438	0.778	0.464	0.842
6	0.486	0.502	****	0.389	****
7	0.651	0.435	0.610	0.474	0.842
8	0.542	0.474	0.667	0.470	0.800
9	0.583	0.445	0.656	0.681	0.919
10	0.579	0.562	0.706	0.627	0.947
11	0.489	0.427	0.675	0.647	0.846
12	0.531	0.478	0.718	0.514	0.842
13	0.675	0.652	0.550	0.452	0.688
Mean	**0.564**	**0.486**	**0.647**	**0.517**	**0.837**

Table A.5: Sub-Forums' Stats

Month	**SF 2**	**SF 3**	**SF 4**	**SF 5**	**SF 6**
1	(45,25,103)	(49,43,145)	(32,40,115)	(60,37,164)	(14,11,49)
2	(19,10,51)	(46,29,169)	(25,8,81)	(47,27,131)	(7,5,13)
3	(35,20,83)	(51,46,252)	(20,13,60)	(58,30,182)	(16,6,33)
4	(38,27,133)	(53,43,196)	(22,15,50)	(36,23,84)	(6,5,13)
5	(32,22,55)	(51,44,184)	(12,8,23)	(55,28,145)	(11,9,30)
6	(33,22,94)	(52,38,208)	(5,3,7)	(53,36,202)	(11,5,13)
7	(26,14,57)	(49,32,173)	(19,10,46)	(55,35,176)	(10,7,52)
8	(38,24,127)	(42,37,171)	(21,17,57)	(45,29,116)	(9,3,13)
9	(35,17,94)	(43,33,174)	(19,10,52)	(25,19,72)	(11,7,41)
10	(35,23,110)	(44,29,138)	(20,9,30)	(34,25,66)	(15,5,27)
11	(38,22,121)	(43,24,124)	(22,9,72)	(25,13,41)	(8,5,37)
12	(31,19,94)	(49,38,156)	(12,8,33)	(42,25,105)	(15,6,36)
13	(27,14,59)	(31,30,102)	(28,17,104)	(38,24,98)	(11,6,27)
Total	**(168,221,1181)**	**(174,351,2192)**	**(96,134,730)**	**(171,282,1582)**	**(501,47,384)**