*Diploma Thesis*

# Mining patterns in non-static, unbalanced, multi-attributed graphs.

Εξόρυξη προτύπων σε μη στατικά, μη ισοζυγισμένα γραφήματα, πολλαπλών χαρακτηριστικών σε κόμβους/ακμές.


*By*

Eirini Milioti

University of Thessaly

School of Engineering

Department of Electrical and Computer Engineering





*Supervisory Commitee:*

Katsaros Dimitrios, Assistant Professor

Bozanis Panayiotis, Professor

Volos, 2017

# ABSTRACT

Due to the connected structure of modern society, networks seem to appear in a tremendous number of topics. Social networks, which represent the collections of social ties among entities have attracted steadily grown interest in the past years. Understanding any piece of information provided, demands a great understanding of small characteristics along with the way that nodes behave and interact with each other. Mining statistically and topologically interesting patterns is an important task within the domain of data mining as it can provide that kind of insight. This task focuses on the identification of frequent subgraphs within graph data sets, the mine of substructures that satisfy certain requests and the supply of a set of estimations regarding the way the network will look in the future.

Social networks are rarely static. Their graph representations evolve as new entities and links are added or deleted all the time. Additionally, they are rarely simple. That means that only one type of relation is not enough to precisely represent and portray them. Even the time that a link appeared on the network constitutes an edge attribute. Issues regarding social network analysis are extensively studied [*31, 32*]. However, the majority of these studies, engage with simple networks, discarding important information. This thesis intends to apply several mining techniques in signed, dynamic graphs, that allow the existence of multiple-type edges between two nodes

and see how the extra information provided by the graph enriches the information mined from the network and helps us predict future possible configurations.

# ΠΕΡΙΛΗΨΗ

Λόγω της διασυνδεδεμένης δομής που εμφανίζει η σύγχρονη κοινωνία, τα δίκτυα φαίνεται να εμφανίζονται σε ένα τεράστιο φάσμα επιστημονικών, αλλά και καθημερινών θεμάτων. Τα κοινωνικά δίκτυα, τα οποία περιγράφουν τις συλλογές κοινωνικών δεσμών μεταξύ οντοτήτων, έχουν προσελκύσει σταθερά αυξανόμενο ενδιαφέρον τα τελευταία χρόνια. Η κατανόηση κάθε πληροφορίας που παρέχεται από αυτά, απαιτεί πλήρη κατανόηση των μικροχαρακτηριστικών του δικτύου, καθώς και του τρόπου με τον οποίο οι κόμβοι αλληλεπιδρούν μεταξύ τους. Η εξόρυξη στατιστικά και τοπολογικά ενδιαφέρουσας γνώσης είναι ένα σημαντικό θέμα στον τομέα της εξόρυξης δεδομένων και μας παρέχει αυτού του είδους την αντίληψη. Η συγκεκριμένη διπλωματική εργασία επικεντρώνεται στην αναγνώριση συχνών υπογραφημάτων στα σύνολα γραφημάτων, στην εξόρυξη δομών που ικανοποιούν συγκεκριμένες απαιτήσεις καθώς και στο να παρέχει πιθανές εκτιμήσεις για τον πιθανό τρόπο εξέλιξης του γραφήματος μελλοντικά.

Η στατικότητα χαρακτηρίζει σπάνια τα κοινωνικά δίκτυα. Τα γραφήματα που τα αναπαραστούν, εξελίσσονται, καθώς νέες οντότητες και νέοι σύνδεσμοι μεταξύ των οντοτήτων εμφανίζονται με την πάροδο του χρόνου. Επιπλέον, τα γραφήματα αυτά είναι σπανίως απλά. Αυτό σημαίνει ότι ένας μόνο τύπος σχέσης μεταξύ των

οντοτήτων (πχ. φιλικός ή επαγγελματικός) δεν αρκεί για να τις απεικονίσει και να τις αναπαραστήσει με ακρίβεια. Ακόμα και η χρονική στιγμή στην οποία εμφανίστηκε μία νέα ακμή στο δίκτυο, αποτελεί χαρακτηριστικό της ακμής. Υπάρχουν πολλές μελέτες σχετικές με το ζήτημα της ανάλυσης των κοινωνικών δικτύων [*31, 32*]. Ωστόσο, οι περισσότερες από αυτές, επικεντρώνονται στην ανάλυση απλών δικτύων, απορρίπτοντας μεγάλης σημασίας πληροφορίες που παρέχει η συνύπαρξη διαφορετικών τύπων ακμών. Στα πλαίσια αυτής της διπλωματικής εργασίας, σκοπεύουμε να εφαρμόσουμε διάφορες τεχνικές εξόρυξης γνώσης σε δυναμικά γραφήματα, στα οποία οι σχέσεις μεταξύ των κόμβων είναι είτε θετικές είτε αρνητικές και επιπλέον επιτρέπουν την ύπαρξη ακμών πολλαπλών τύπων μεταξύ δύο κόμβων, και να δούμε πώς οι επιπλέον πληροφορίες που παρέχονται από το γράφημα εμπλουτίζουν τις εξαχθείσες πληροφορίες και μας βοηθούν να προβλέψουμε και να προσεγγίσουμε το πώς θα διαμορφωθεί μελλοντικά το δίκτυο.

# ACKNOWLEDGEMENTS

I offer my sincerest gratitude to my supervisor, Dr Dimitrios Katsaros, for giving me the opportunity to work on this project and providing me with all the support, guidance and encouragement required to carry it out. I also want to thank my co-advisor, Dr Panayiotis Bozanis. Furthermore, to all my friends that have helped me, I want to say thank you. Last but not least, I want to express my deep gratitude to my family, who have supported my efforts in order to achieve my every goal, throughout the years of my studies.

# CONTENTS


Abstract . . . . . iii

Περίληψη . . . . . v

Acknowledgements . . . . . vii

List Of Figures . . . . . x

List Of Tables . . . . . xii

1. Introduction To Graphs and Patterns . . . . . 1

1.1 Introduction . . . . . 1

1.2 Graphs . . . . . 2

1.2.1 Graphs that change over time . . . . . 2

1.2.2 Graphs with multiple edge attributes . . . . . 3

1.2.3 Signed Graphs . . . . . 5

1.2.3.1 Structural Balance . . . . . 5

1.3 Graph and Link Patterns . . . . . 7

1.3.1 Maximal Frequent Subgraphs . . . . . 8

1.3.2 Link Type Prediction . . . . . 10

1.3.3 Predicting Link Existence . . . . . 10

1.3.4 Pattern mining with restrictions . . . . . 11

2. Algorithms . . . . . 12

2.1 Mining Interesting Patterns in Signed, Attributed Graphs; Description of the Problem . . . . . 12

2.2 Basics, Algorithms and Examples . . . . . 13
2.2.1 Pattern mining with restrictions . . . . . 15
2.2.1.1 Algorithm1;Listing Triangles and Examples . . . . 15
2.2.2 Subgraph Mining . . . . . 18
2.2.2.1 Algorithm2;Maximal Subgraph Mining . . . . . 18
2.2.3 Link Prediction . . . . . 20
2.2.3.1 Algorithm3;Likelihood of a link to appear . . . . 20
2.2.3.2 Link Sign Prediction . . . . . 22
3. Data and Results . . . . . 24
3.1 Description of the Datasets and Results . . . . . 24
3.1.1 Dutch College Freshmen . . . . . 25
3.1.2 SG&R Law Firm . . . . . 31
3.1.3 Slashdot . . . . . 36
4. Conclusions and Discussion . . . . . 40
References . . . . . 42

# LIST OF FIGURES


**Figure 1** Florentine families' social network

**Figure 2** Signed Network that depicts the evolution of alliances in Europe from 1872 to 1907

**Figure 3** Configurations of balanced triangles

**Figure 4** Configurations of unbalanced triangles

**Figure 5** Othello's social network

**Figure 6** Examples of biological graphs

**Figure 7** An example of a signed network

**Figure 8** An example of a network with multiple types of edges

**Figure 9** Evaluation of the probability that two nodes of the Dutch College Freshmen network will connect in the future, as a function of the number of common neighbors they have

**Figure 10** Probability of an edge to appear and join a pair of nodes of the Dutch College Freshmen network in the near future, based on our approach. The predictor had as input a limited image of the network

**Figure 11** Probability of an edge to appear and join a pair of nodes of the Dutch College Freshmen network in the near future, based on our approach

**Figure 12** Probability of an edge to appear and join a pair of nodes of the Dutch College Freshmen network in the near future, based on our approach. The prediction for a currently missing edge, is based on all three snapshots of the network

**Figure 13** Possible configurations of triangle patterns between nodes that are interconnected with edges of the same type in the SG&R Law Firm network

**Figure 14** Possible configurations of triangle patterns between nodes that are interconnected with edges of different types in the SG&R Law Firm network

**Figure 15** Evaluation of the probability that two attorneys of the SG&R Law Firm network will connect in the future, as a function of the number of common neighbors they have

**Figure 16** Evaluation of the probability that two nodes of the Slashdot network will connect in the future, as a function of the number of common neighbors they have

# LIST OF TABLES

**Table 1** List of notations used in the algorithms and their interpretation
**Table 2** Datasets and their characteristics
**Table 3** Number and types of triads in Dutch College Freshmen network
**Table 4** Number and types of triads in SG&R Law Firm network
**Table 5** Triangle pattern interconnections between 774 nodes in Slashdot network
**Table 6** Triangle pattern interconnections between 7736 nodes in Slashdot network

# Chapter 1

# INTRODUCTION TO GRAPHS AND PATTERNS

## 1.1 Introduction

This chapter will provide an introduction to graphs and pattern mining. We live in a connected world. Our everyday lives are full of complex systems, each consisting of entities and relations between them. For example, when someone tells a story to someone else, it is extremely likely that this story will travel to other people as well, forming a network of interactions. Graphs are widely used to represent such relations among objects, in order to be used in applications, such as web analysis, computer vision, video indexing, social networks, bioinformatics, chemical and text retrieval. Entities in the data are represented by nodes, while the relations between them are represented by edges that connect the nodes. An email network's graph for instance, would have email accounts as nodes and email exchanges as edges, while a protein-protein interaction network's graph would have labeled proteins as nodes and the interactions between them as edges. Given a dataset, a topic of interest is to discover interesting patterns. The main goal is to extract statistically significant and useful knowledge from the given data [*15*]. Structured and semi-structured data can be

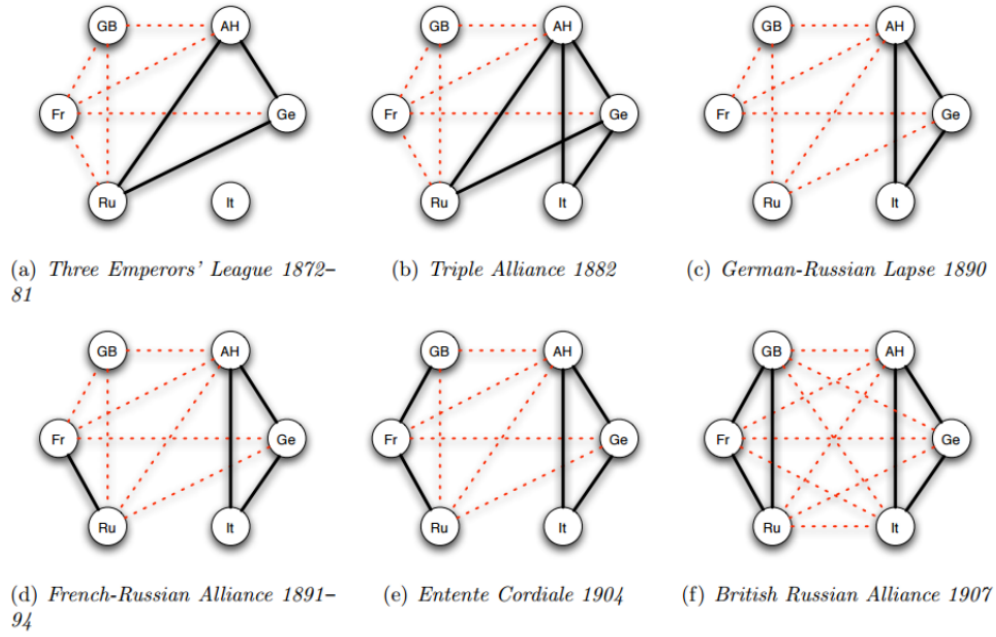
represented easily by graphs. Due to this fact, there exists increasing interest in the mine of graph data. There are many different types of graphs, but we are particularly interested in the problem of mining patterns and finding statistically interesting behaviors in dynamic graphs whose edges have multiple attributes and the nodes are connected with positive or negative links, denoting the relationship between them.

## 1.2 Graphs

As we previously mentioned, graphs provide a natural way of representing connected entities, appearing whenever it is useful to represent how things are either physically or logically linked to one another in a network structure. A graph is often denoted by $G = (V, E)$, where $V$ is the set of nodes and $E$ is the set of edges.

### 1.2.1 Graphs that change over time

In real-world systems relations between objects are rarely static. Their graph representations evolve as relations are created or stop existing over time. For example, proteins interact from time to time, messages on social networks are sent at a certain point, new friendships are created constantly etc. A time evolving graph contains a sequence of static graphs $\{G_1, G_2, \ldots, G_n\}$, where $G_t = (V_t, E_t)$ is a snapshot of the evolving graph at a timestamp $t$. Each link appears at a specific point in time. As shown in Figure 2, dynamic graphs are represented by a series of static snapshots taken at various points in time. Possible modifications consist of new nodes and edges being added to the network, the attributes of the already present edges and nodes being modified or previously present edges being removed from the network.




(a) *Three Emperors' League 1872–81* (b) *Triple Alliance 1882* (c) *German-Russian Lapse 1890*

(d) *French-Russian Alliance 1891–94* (e) *Entente Cordiale 1904* (f) *British Russian Alliance 1907*

*Figure 2: The evolution of alliances in Europe, 1872-1907(the nations GB, Fr, Ru, It, Ge and AH are Great Britain, France, Russia, Italy, Germany, and Austria-Hungary respectively). Solid dark edges indicate friendship, while dotted red edges indicate enmity. Note how the network slides into a balanced labeling, and eventually into World War I. This figure and example are from Antal, Krapivsky, and Redner [9].*

### 1.2.2 Graphs with multiple edge attributes

In many real world problems, connections between entities are defined by several different relationships. For example, similarity between two scientific articles can be defined based on authors, citations to, citations from, keywords, titles, where they are published, text similarity, etc. Social interactions between a group of individuals can be based on the nature of each relationship like business, family, friendships, or the means of communication like phonecalls, instant messages, emails, personal meetings. Electronic files can be grouped by their type, name, the time they were created, or the pattern they are usually accessed. In these examples, there are multiple graphs that define relationships between the subjects [*13*].

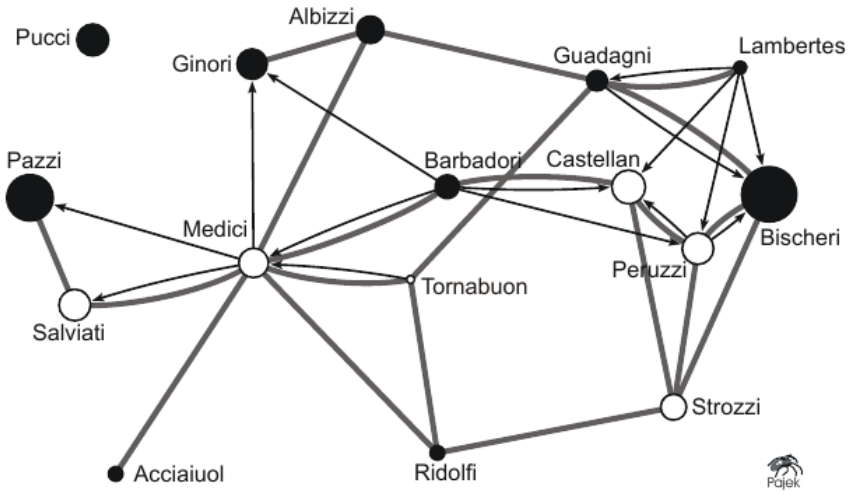



*Figure 1: Social Graph of marriage and business relations between 16 Florentine families in 1400 AD*

For instance, the graph shown in Figure 1 illustrates the relations of marriage and business among 16 families in Florence around 1400 AD. The data are part of a larger dataset collected and analyzed by John F. Padgett and C. K. Ansell. Each family is represented by a circle. Directed lines (black edges) represent business relations among families, pointing towards the more prosperous family. Marriage relations between families are undirected lines (gray edges) [*14*]. Padget and Ansell explain in their work [*18*] the way that the Medici became an economic force and great political influence via establishing their family through the marriage network, as more than half the paths relating the 16 families pass through them.

### 1.2.3 Signed Graphs

Relations between nodes can often reflect positive (1), negative (-1) or neutral (0 or absence of an edge) interaction. The type of the relationship affects the structure of the network. Positive edges denote friends, fans, followers or collaborators, while negative links denote foes, disagreements, controversy or antagonism. Signed graphs can be used for modeling interactions in chemical or biological networks, in social network analysis, communication networks, power systems, sociometric structures or to represent political and economical relations. For example, users on Wikipedia can vote for or against the nomination of others as admins [*1*], participants on Slashdot can declare others to be either friends or foes [*2*], links between blogposts of different bloggers can be positive when the one blogger endorses the statements of the other or negative if the users express difference in opinions [*3*]. Signed graphs provide patterns of interaction [*4*].

#### 1.2.3.1 Structural Balance

If we consider a small world social network where everyone knows everyone (e.g. a classroom or an office), or the setting for international relations between countries and their diplomatic position towards one another (Figure 2), a question that needs to be answered is whether or not this network is balanced.

Structural balance is based on theories in social psychology dating back to the work of Heider in the mid-20th-century [*5*], and generalized and extended to graphs beginning by Cartwright and Harary's work in the 1950s [*19*]. The key idea is that every two people in the network will be either friends or enemies. If we extend the idea to three people, then we have a combination of positive and negative relations. Some combinations are more likely and socially accepted than others. Particularly, there are four possible combinations of relations among these nodes. Firstly, there can exist three pluses {+, +, +} among them, which is a very common and normal situation, that indicates mutual friends. Another possibility is that of two minuses and a plus {-, -, +}, which is also a rather common case, that indicates two friends who

have a common enemy. A different case is having a triangle with two pluses and a minus {+, +, -}, which corresponds to a complicated situation, where a person has a friendly relationship with two individuals, who do not get along with each other. Lastly, a triangle with three minuses {-, -, -} is also a complicated and unusual possibility, where we have three mutual enemies. In the last two combinations there are usually forces that tend to uncomplicate and balance the situation. For further understanding, lets say that we have three people, Alice, Bob and Charlie, where Alice is friends with the other two, but Bob and Charlie have unfriendly feelings for each other. Then there would exist two possibilities. In the first one, Alice would try to persuade the others into becoming friends so the triangle would convert to three pluses {+, +, +}, or one of the them would probably persuade Alice to side with him, so the triangle would convert to a plus and two minuses {+, -, -}. On the other hand, if Alice, Bob and Charlie had all hostile feelings for each other, two of them would probably team up against the other. Based on the theory that in the first two cases there are no forces trying to change the situation, in contrast to the last two, we refer in such stable triangles as balanced (Figure 3), while we use the word unbalanced (Figure 4) for the unstable ones.

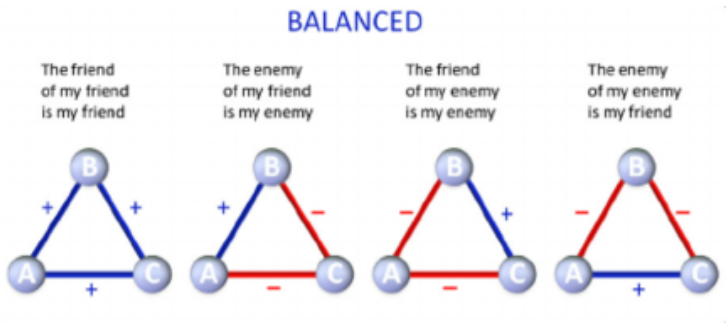



*Figure 3: Balanced Triangles*

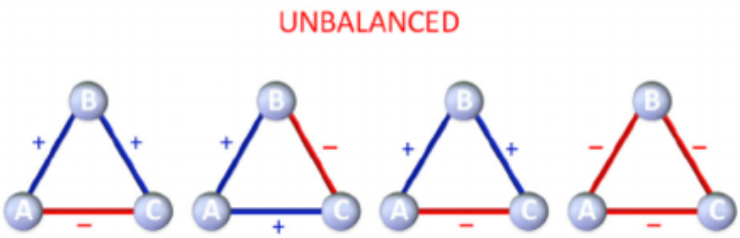



*Figure 4: Unbalanced Triangles*

Structural balance theory has been developed extensively in the time since [*10*], including the work of Davis, who studied the formulation of a variant proposed as a way to eliminate the assumption that "the enemy of my enemy is my friend" [*11*]. In particular, weak structural balance posits that only triangles with exactly two positive edges are implausible in real networks and that all other kinds of triangles should be permissible.

## 1.3 Graph and Link Patterns

The discovery of patterns in graph data provides an insight of useful knowledge and meaningful information for many applications, via the embodiment of descriptive and predictive modeling, including compact representation of the information, creating friend suggestions in social networks and finding frequent molecular structures in biological networks. By mining patterns, we are able to answer questions like how many triangles exist in a graph, which subgraphs are frequent, what is the possibility of an edge to appear in the future, what defines a normal or an abnormal behavior for the graph, thing that may indicate fraud or spam, user behavior predictions, user preference based applications etc. Or in a more comprehensive way, questions such as what products are often purchased together, in which products are users interested in based on their model user profile, what reaction will certain organisms have in a certain drug based on their DNA, what is the role of a user in the network based on his relations with other users , who is the leader or the main character of a network (Figure 5), who should someone follow in order to see the most interesting or popular content etc. We can also use frequent subgraphs to monitor the health of a network as it evolves, which is an important tool for a cyber-security analyst who is monitoring a large cyber-network and may be alerted to a potential attack due to the change in the frequent subgraphs.

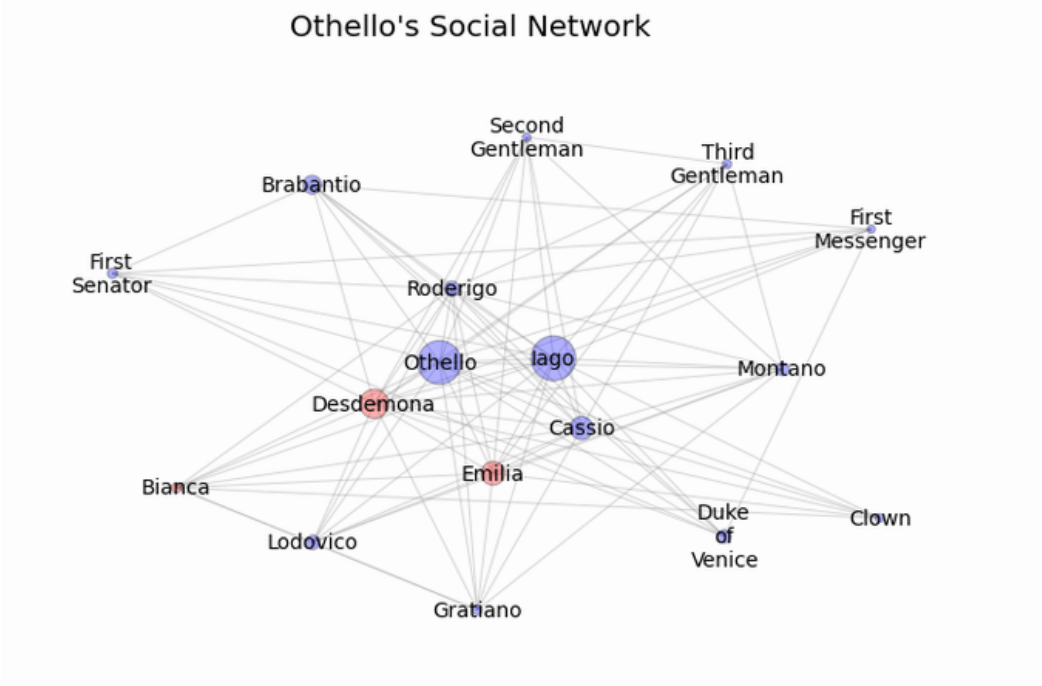



*Figure 5: Othello's Social Network. Two characters are connected if they appear in the same scene together. The color of the nodes is assigned by gender and their size is based on the number of lines the node is part of. The bigger the size of the node, the more socially central the character is in the play. By this model, we can easily see that Desdemona, Iago and Othello appear to be the most crucial characters. [12]*

### 1.3.1 Maximal Frequent Subgraphs

Among the various kinds of patterns that someone can discover in a graph, frequent subgraphs are the most basic. The problem of frequent subgraph mining is to find any subgraph *g*, whose occurrence counts are above a predefined threshold across a set of graphs (Figure 6). The number of possible frequent subgraphs increases exponentially with the size of the graph i.e. for a frequent k-graph, the number of its frequent subgraphs can be as large as 2k [*16*].

As the set of maximal frequent subgraphs is much smaller compared to the set of frequent subgraphs, maximal frequent subgraphs are proposed as mechanisms to limit the number of frequent subgraphs generated and provide a more meaningful set by encoding the maximal common structures in a set of graphs. According to studies [6, 7, 8], in the case of biological networks, protein contact maps and metabolic pathways, they appear to be the most interesting patterns.

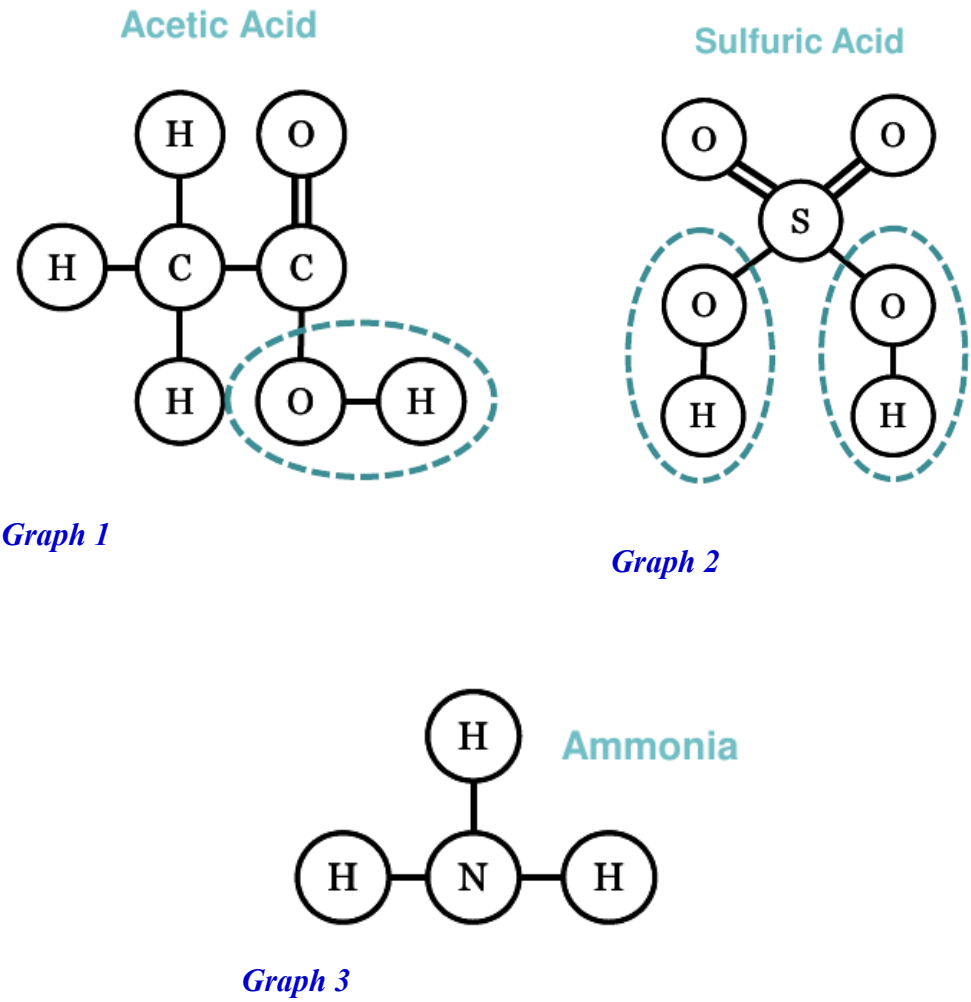



*Figure 6: Common occurrences of hydroxide-ion, which, as we can see, exists in two of the graphs of the dataset (Graph 1 and Graph 2), while it is absent from the third one (Graph 3).*

### 1.3.2 Link Type Prediction

In dynamic networks, where new edges appear, indicating new interactions between objects, it is useful to predict links. That is, given a snapshot of a network at time *t,* we wish to predict the edges, that will appear in the network during the interval from time *t* to a future time, *t',* along with their types.

By predicting the type of the relationships between objects, based on the properties of the objects involved, more information is made available to the mining process. Given epidemiological data, for instance, we may try to predict whether two people who know each other are family members, coworkers, or acquaintances, by using the attributes of the linked entities. In another example, we may want to predict whether there is an advisor-advisee relationship between two coauthors. Given Web page data, we can try to predict whether a link on a page is an advertising link or a navigational link. [*20*]

### 1.3.3 Predicting Link Existence

Unlike link type prediction, where we know that a connection exists between two objects, we now want to predict the existence of a link between them. In order to do that, we must consider links that exist and links that do not. Examples include predicting whether there will be a link between two Web pages, or given a social network of coautorship among scientists, we try to predict whether or not, two scientists that are close in the network, may collaborate in the future. In epidemiology, we can try to predict with whom a patient came in contact. [*20*]

### 1.3.4 Pattern mining with restrictions

This type of mining aims to identify all frequent patterns displaying a specified constraint, for instance finding all triangles on a graph, all patterns that have a certain degree, every node that has only positive or negative relationships etc.

# Chapter 2

# ALGORITHMS

## 2.1 Mining Interesting Patterns in Signed, Attributed Graphs; Description of the Problem

In recent years there has been an explosion in the amount of data available for network analysis leading to various types of networks, forcing us to be creative with the description of graphs in order to fit all their particular features. Network data involving relational structures representing interactions between entities, are commonly represented by graphs. Analyzing the data that are derived from such networks results to much gained information.

Studies in social network analysis are mainly focused on graphs consisting of binary relations between the nodes, i.e. a line between a pair of entities either exists or not, considering almost exclusively one type of positive relationship. We find that this approach discards essential information attached to the original network. For instance, consider a network with nodes representing different students of a school. The edges

then may indicate coexistence in the same classroom, family ties, cooperation, friendship, antagonism or dislike. In order to understand the structure and underlying social mechanisms of the network, these different types of edges should be considered in our analysis. Therefore, in our approach, in order to provide a fuller insight into patterns of interactions, we will use techniques to analyze and mine the structure of such graphs and see how this structure affects the evolution of the network.

With this in mind, our first task is to find subgraphs that best match our interest (e.g. to find all triangles connected with same-type-edges) in a graph. Next, we identify maximal frequent subgraphs. Lastly, we will adapt the structural balance theory to examine the interplay between friendly and non-friendly links in the network and recognize the differences between the true and predicted configurations of the various types of links.

## 2.2 Basics, Algorithms and Examples

For the rest of the thesis we assume that our dataset is a set of different instances of a social graph at different times that each has a static set of nodes and a dynamic set of attributed edges. The edges of the graph are undirected and signed. Each one has a certain connection type and in the case that a line or more exist between two entities, there will be an assigned weight (1 for a positive sign, or -1 for a negative sign) that denotes the feeling of the relationship. Multiple edges connecting the same two nodes with different connection types, are allowed. The graphs are simple, that means that they do not contain loops( $(i,i)$ edges do not exist). For each edge, we know the time of its first appearance on the graph. Not all edges need to be present; a non-existing edge between two nodes corresponds to neutral or absent feelings. There are four possible configurations of signed triangles. In the theory of social balance, the triangle configurations with an even number of minus signs are considered as balanced. On the other hand, the two configurations with an odd number of minus signs correspond to unbalanced triangles. So if a triangle has a product of weight greater than 0, it is

balanced, otherwise, it is considered unbalanced. Table 1 lists the notations used below.

| Notation | Interpretation |
|---|---|
| $G = (V, E)$ | The data graph with vertex set $V$ & edge set $E$ |
| $R = \{r_1, r_2, \ldots, r_l\}$ | The set of $l$ attributes defined on $V$ |
| $Si$ | The edge array containing the sign of each edge. 1 for positive sign, -1 for negative sign. |
| $A$ | The edge **x** attribute matrix denoting the type of the edge |
| $G_t = (V, E, t)$ | Snapshot of the graph at time t |
| $N(V)$ | The set of nodes that share a link with each node $v$, signifying the neighbors of $v$ |
| $C = \{e_1, e_2, \ldots, e_k\}$ | Array that contains the labels of the $k$ edges included in graph $g$ |
| $Smax$ | The maximal frequent subgraph of the dataset |
| $support$ | The number of snapshots that contain the pattern |
| $minsup$ | The support threshold |
| $code(g)$, $C_i$ | Array that contains the edge list of graph $g$ |
| $Code$, $C$ | Array that contains the common edge list between a set of graphs |

*Table 1. Notations used in the algorithms below.*

### 2.2.1 Pattern mining with restrictions

We, firstly, address the problem of finding subgraphs (triangle patterns in particular) that best match our needs. A typical constraint specifies both connectivity patterns between nodes and the type of their connection. For example, in the above school network, a constraint may have a triangle pattern where every node like each other and they have also participated in a group project together. In a problem like that, we have to consider structure, attributes and signs to exact match our demands.

The reason why we present a high interest in triangle patterns is, that counting the number of triangles in a graph has gained importance over the last years, since several significant graph mining applications rely on computing them in the graph of interest. In our case, it is significant to find the triangles in the dataset, in order to determine whether or not the graph and each node are balanced. Other metrics that involve the execution of a triangle counting algorithm in order to be computed, are the clustering coefficient and the transitivity ratio. Recently, in [*22*] it was shown that triangles can be used to detect spamming activity. Moreover, [*23*] shows how triangles can be used to uncover the hidden thematic structure of the web.

#### 2.2.1.1 Algorithm1;Listing Triangles and Examples

To begin with, we will present a way to count and list all triangles present in our graphs, along with their balance situation and their type configuration. The key idea of our algorithm, is that only the lowest alphabetically vertex of a triangle is responsible for counting it. Furthermore, it is very useful and applicable in all possible scenarios.

| **Algorithm1** Listing Triangles |
|---|
| **Input:** graph $G = \{V, E, N(V), Si(E)\}$, where $V$ is the set of the $N$ nodes of the graph, $E$ is the set of e edges, $N(V)$ is the set of neighbors of every node $v \in V$ and $Si(E)$ is the list of signs corresponding to every edge in $E$. |
| 1. **foreach** vertex v ∈ V **do**<br>2. **foreach** $u \in N(v)$, so that $u$ is higher alphabetically than $v$ **do**<br>3. **if** there exists a vertex $w$, so that $w \in N(v)$ **&** $w \in N(u)$, and in the same time $w$ is higher alphabetically than $v$ and $u$, **then** $u,v,w$ form a triangle |
| **Output:** list of existing triangles |

Considering the case that we want to find, for instance, only the balanced triangles that exist in the graph, we will modify the algorithm in order to fulfill our needs. In the case of the balanced triangles, we should check if the sum of the edge signs in the triangle equals 3 (case of mutual friends) or -1 (case of two friends sharing a mutual enemy).

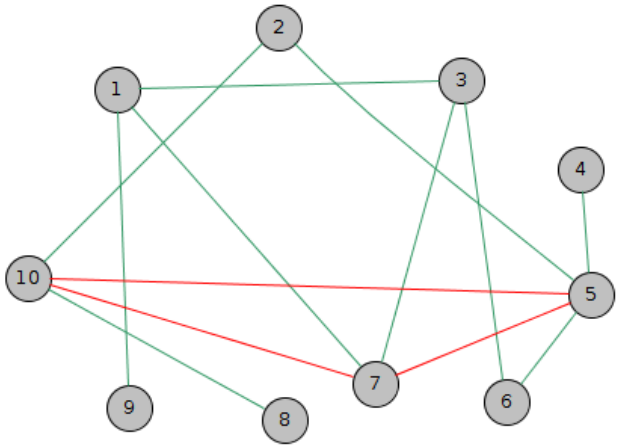



*Figure 7. The first 10 nodes of a data set collected among a group of university freshmen. Green edges indicate friends or friendly relations, while red edges indicate troubled relationships. [21]*

The data in Figure 7 were collected among a group of university freshmen who, except for a few existing relationships (acquaintances from a former school), did not know each other at the first measurement. The students were asked to rate their relationships [23]. Applying *Algorithm1* in the signed network of Figure 7, would return three triangles ({*1, 3, 7*}, {*2, 5, 10*}, {*5, 7, 10*}), two unbalanced ({*2, 5, 10*}, {*5, 7, 10*}) and a balanced {*1, 3, 7*}.

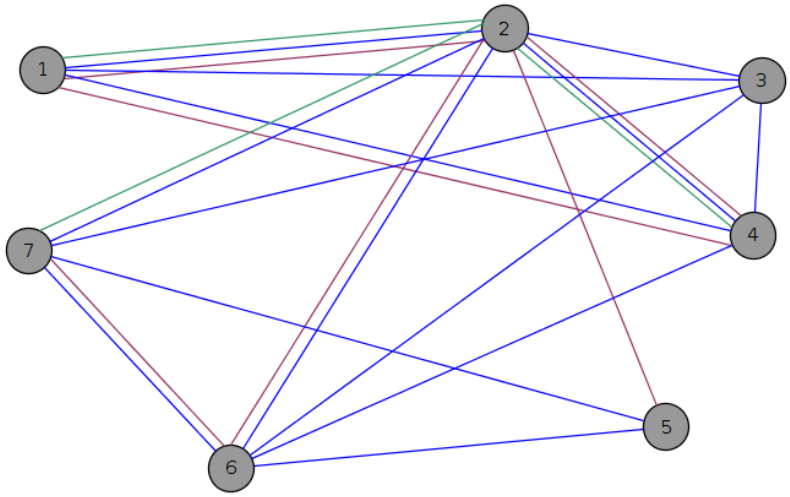



*Figure 8. The first 7 nodes of a data set collected from the managers of a high-tech company. Each manager was asked "To whom do you go to for advice?" (blue edges) and "Who is your friend?" (dark red edges).Data for the item "To whom do you report?" (green edges) were taken from company documents. [33]*

The data in Figure 8 were collected from a high-tech company's managers. The company manufactured high-tech equipment on the west coast of the United States and had just over 100 employees with 21 managers. Each manager was asked "To whom do you go to for advice?" and "Who is your friend?". Data for the item "To whom do you report?" were taken from company documents.

### 2.2.2 Subgraph Mining

The discovery of graph structures that occur a significant number of times across different snapshots of a dynamic graph can unveil new and useful knowledge hidden in the dataset. For instance, we can detect stable and regular sets of node relationships or which are the most influential communities among the graph. In our case, we are not only interested in extracting topological patterns that frequently exist among nodes, but also in the type of the relation that connects said nodes.

Using only maximal frequent subgraphs instead of using all the patterns is one of the techniques that aim to avoid redundancy among the computed patterns and consequently reduce the dimensionality of the set. A maximal frequent subgraph is a pattern which is not a subgraph of any other frequent subgraph [*24*]. From the frequent maximal subgraphs it is possible to reconstruct the whole set of frequent subgraphs, as they are all summarized into the maximal patterns.

#### 2.2.2.1 Algorithm2;Maximal Subgraph Mining

Given a set of graphs *G*, the support of a subgraph *s* is defined as the fraction of graphs in *G* in which *s* occurs [*30, 29*]. *s* is frequent if its support is at least a user specified threshold, which is referred to as the support threshold. Given our dataset of graph snapshots, we define the support as the number of graph snapshots that *s* occurs and set the support threshold to at least *No of snapshots-1*.

| **Algorithm2** Maximal Subgraph Mining |
| --- |
| **Input:** codes $C_1$, $C_2$, ..., $C_n$ containing the edge list of the $g_1$, $g_2$, ..., $g_n$ snapshots respectively, *minsup,*which is the support threshold, set to numberOfGraphSnapshots-1 |
| 1. **find** the global code $C = \{e_1, e_2, \ldots, e_k\}$ that contains the $k$ frequent edges of the dataset, meaning the edges that exist in at least *minsup* codes<br>2. **choose** edge $e_i$ , **insert** $e_i$ in $S$ and **delete** $e_i$ from $C$<br>3. **do**<br>4. **find** edge $e_j$ so that $e_j$ contains one of the nodes that create $e_i$ (In a Breadth first manner) and **insert** $e_j$ in $S$<br>**until** there are no remaining edges in C, or the remaining edges do not have common nodes with the edges already chosen<br>5. **if** $C$ is not empty, **goTo** step 2 and create a new Subgraph $S$ |
| **Output:** *Smax,* maximal frequent subgraph set in the dataset |

In a similar manner as before, we can modify the input array in order to include only the type of edges that we wish the output subgraphs to contain. Another modification to the algorithm would be to ignore certain differences. For instance, in the above example of a school network, the switch of an edge type from cooperation to friendship would not cause that much of an impact in the network, as would the swift from friendship or cooperation to antagonism. It all depends on what we are looking for.

### 2.2.3 Link Prediction

The attitude of one user toward another can be estimated from evidence provided by their relationships with other members of the surrounding social network [*25*]. Additionally, being part of a particular community, business, school, voluntary organization or neighborhood, being a patron in a certain place, being a member in a certain gym, are all activities that, when shared between two people, tend to increase the likelihood that they will interact and hence form a link in a social network [*28*]. Naturally, if two people are friends, they influence each others friendships and enmities.

#### 2.2.3.1 Algorithm3;Likelihood of a link to appear

In a social network, people tend to create new relationships with people that are closer to them. Based on the dynamic nature of our graph, we will empirically estimate the possibility of a connection to appear between two nodes, by taking as metric the number of common neighbors they had in two different snapshots of the network. A link is significantly more likely to be positive when its two endpoints have multiple neighbors (of either sign) in common. This observation is consistent with qualitative notions of social capital [*27*], as users with common neighbors have relations that are in plain site for others to see in a social sense, and hence have greater implicit pressure to remain positive.

**Algorithm3** Likelihood of a link to appear

**Input:** two snapshots of the graph at times $t_1$ and $t_2$, where **g**raph $G_{t1} = \{V, E_1, N_1(V), S(E_1)\}$, where $V$ is the set of the $N$ nodes of the graph, $E_1$ is the set of $e_1$ edges, $N_1(V)$ is the set of neighbors of every node $v \in V$ and $S(E_1)$ is the list of signs corresponding to every edge in $E_1$ and graph $G_{t2} = \{V, E_2, N_2(V), S(E_2)\}$ respectively.

1. **foreach** node $v \in V$ **do**
2. **foreach** neighbor $n \in N_1(v)$ do
3. **foreach** node $w \in V$, so that $w \neq v$ and $w \notin N_1(v)$ **do**
4. **foreach** neighbor $b \in N_1(w)$
5. **if** $n == b$ **then**
6. **increase** the number of common neighbors between $v$ and $w$
7. **for** $k=1$ to $max$(number of common neighbors between two nodes) **do**
8. **foreach** node $v \in V$ **do**
9. **foreach** node $w \in V$, so that $w \neq v$ and $w \notin N_1(v)$ **do**
10. **if** (the number of common neighbors between $v$ and $w == k$) **then**
11. **if** $w \in N_2(v)$ **then**
12. **increase** the number of unlinked nodes of $G_1$ that connected in $G_2$
13. **else increase** the number of unlinked nodes of $G_1$ that did not connect in $G_2$
14. $T(k)$ = the fraction of the pairs that have formed an edge by the time of the second snapshot

**Output:** $T(k)$, an empirical estimation of the probability that two people will form a link in the future, as a function of the number of common friends they have

#### 2.2.3.2 Link Sign Prediction

A more useful estimation for the considered signed networks would be to find the possible sign that the certain link will have, given the information provided by the network so far. Signed link prediction has connections to social balance theory while no such connection exists for the unsigned link prediction task. In order to do that, we will modify our algorithm as follows. In addition to the evaluation of the common neighbors between two nodes, we will also consider the negative or positive relation between each neighbor with both of the nodes.

The basic idea is that the sign of an edge should minimize the number of unbalanced triangles involving this particular edge. So for each choice of the sign of an edge, we choose the sign that causes it to participate in a greater number of triangles that are consistent with balance theory. The logic of social balance theory indicates that "the enemy of my friend is my enemy," "the friend of my enemy is my enemy," and variations on these [*26*]. Consider the situation in which a user A links positively to a user B, and B in turn links positively to a user C. If C then forms a link to A, what sign should we expect this link to have? Balance theory predicts that since C is a friend of A's friend, so we should see a positive link from C to A. So based on the ideas that if w forms a triad with the edge *(u, v)*, then structural balance theory posits that *(u, v)* should have the sign that causes the triangle on *{u, v, w}* to have an odd number of positive signs, we come up with this possibilities.

- *Case 1:* If both node A and node B have balanced relations with the majority of their common neighbors, then if the link between them will appear, it will probably be a positive one, as the friends of your friends are also your friends.
- *Case 2:* If both node A and node B have unbalanced relations with the majority of their common neighbors, then if the link between them will appear, it will also be a positive one, based on the idea that the enemies of your enemies are your friends.

- *Case 3:* If node A (respectively node B) has unbalanced relations with the majority of common neighbors, while B (respectively node A) has mainly balanced relations with the majority of them, then if the link between them will appear, it will probably be a negative one, based on the idea that the enemies of your enemies are your friends.

# Chapter 3

## DATA AND RESULTS

### 3.1 Description of the Datasets and Results

In this chapter, we will apply mining techniques in dynamic social networks, with multiple types of edges. Our main goal is to determine whether or not the existence of multiple and/or signed edges, helps us in the mining process.

Table 2 lists the real world datasets we used in our experiments in order to evaluate the extra information we receive, when we allow different types of ties with both negative and positive signs, to exist in the network. We have some smaller datasets with hundreds of edges and some larger ones with thousands of edges.

| Datasets | Number of Snapshots | Static Number Of Nodes | Number of Edges | Distinct Edge Attributes |
|---|---|---|---|---|
| **Dutch College Freshmen** | *3* | *32* | *1st Snapshot: 600*<br>*2nd Snapshot: 390*<br>*3rd Snapshot: 424* | *6* |
| **Slashdot1** | *3* | *774* | *1st Snapshot: 3,253*<br>*2nd Snapshot: 3,239*<br>*3rd Snapshot: 3,325* | *2* |
| **Slashdot2** | *3* | *7736* | *1st Snapshot: 105,966*<br>*2nd Snapshot: 105,272*<br>*3rd Snapshot: 106,327* | *2* |
| **SG&R Law Firm** | *3* | *71* | *1st Snapshot: 1,139*<br>*2nd Snapshot: 852*<br>*3rd Snapshot: 465* | *3* |

*Table 2: Datasets and their Characteristics*

### 3.1.1 Dutch College Freshmen

The first dataset, which was collected by Gerhard van de Bunt [*33*], is a relatively small, signed graph that contains the relationships between a group of university freshmen and how they evolved in a period of 24 weeks. At the beginning of the first week, only a few acquaintances between them existed from a former school. Other than these, the students did not know each other. The original dataset was collected at 7 time points. The first four time points were three weeks apart, whereas the last three time points were six weeks apart. The original group consisted of 49 students, but some of them did not continue with their studies, while others did not fill the questionnaire, leading to a group of 32 students, for whom almost complete data are available. The students were asked to rate their relationships on a six point scale, with ratings corresponding to:

- *1-4:* Best friends, friendly relationships and pleasant contacts
- *5:* Troubled relationships and conflicts
- *0, 6, 9:* People who don't know each other or absent data

The dataset also provides information regarding the smoking behavior of the students and their education program. These attributes supply information regarding the contact situation between the students, as smokers had to separate themselves in special areas and students attending different education programs had different schedules and courses, even though they had all started their education at the same time.

We converted the data in graph snapshots. We proceeded to group them in three snapshots, in order to increase the density of each of them. The first snapshot contains the relationships that appeared in the first 9 weeks, the second one contains the next 6 weeks and the last one the last 12 weeks, as the two last time points were more distant. Each node corresponds to a student and each edge denotes a relationship between two students. We assigned signs and attributes (friendship, enmity, smokers, same program) to the edges. If the relationship is rated between 1 and 4, then the sign is positive. If the relationship is rated as 5, then the sign is negative. For every other rating, the edge is absent. The average snapshot size, in terms of number of edges, is 471.3 . The largest snapshot contains 600 edges. The average number of edges connecting two nodes is 1.2 and the maximum number of edges between two nodes is 3. The 81% of the edges denote either friendship or enmity, the 6% denote smoking habits and the last 13% ,the type of study program that connects the entities.

| | Snapshot 1 | Snapshot 2 | Snapshot 3 |
|---|---|---|---|
| **Number of Triangles** | *6824* | *2615* | *3163* |
| **Number of Signed Triangles** | *324* | *298* | *439* |
| **Balanced Triangles** | *321* | *296* | *438* |
| **Unbalanced Triangles** | *3* | *2* | *1* |
| **Type Of Triangle: + + +** | *306* | *286* | *429* |
| **Type Of Triangle: + - -** | *15* | *11* | *9* |
| **Type Of Triangle: + + -** | *2* | *1* | *1* |
| **Type Of Triangle: - - -** | *1* | *1* | *0* |
| **Same-Type-Edges Triangles** | *1300* | *298* | *439* |
| **Different-Type-Edges Triangles** | *4980* | *1748* | *2069* |

*Table 3: Numbers and Types of the first network's triads*

Table 3 gives the counts of the possible triangle configurations of the network. The number of triangles is different from the number of signed triangles, as not all types of edges have signs. The edges that have signs are those that indicate the existence of a friendship or an enmity between two nodes. By considering all types of edges in the network, we extract almost 23 times more triangle patterns, than we would have in the case we considered only the friendship ties.

As we can see, in all three snapshots the all-positive triad is overrepresented, while the triads consisting of two enemies with a common friend, along with the triads of three mutual enemies, are underrepresented. Even though the number of unbalanced triangles in the dataset is relatively small, the structural balance theory is consistent with the data. Unbalanced triangles gave in to the social forces and reversed their signs in order to become balanced. Considering the configurations that balance theory suggests, along with the theory of weak balance, we evaluated the prediction accuracy of the sign of the new edges. The probability of the new edge, having the sign we predicted, was as high as 0.7 . This can also be explained by the notion from social-

capital theory, saying that pressure is exerted on the entities on display in order to remain positive and maintain the harmony of the community.

By systematically going through the relative frequency or proportions of edges within and between vertex pair categories, we can detect conditional dependencies of types of edges. Within the category of social relations (friendship or enmity) , there are 60% of vertex combinations where a social relation exists along with a common smoking condition. Also there are 32.7% vertex combinations where a social relation exists along with a common program attendance. From these results we can conclude, that there is a high tendency to create social relations within the programs of the university. We also detect tendencies to social relations between smokers.

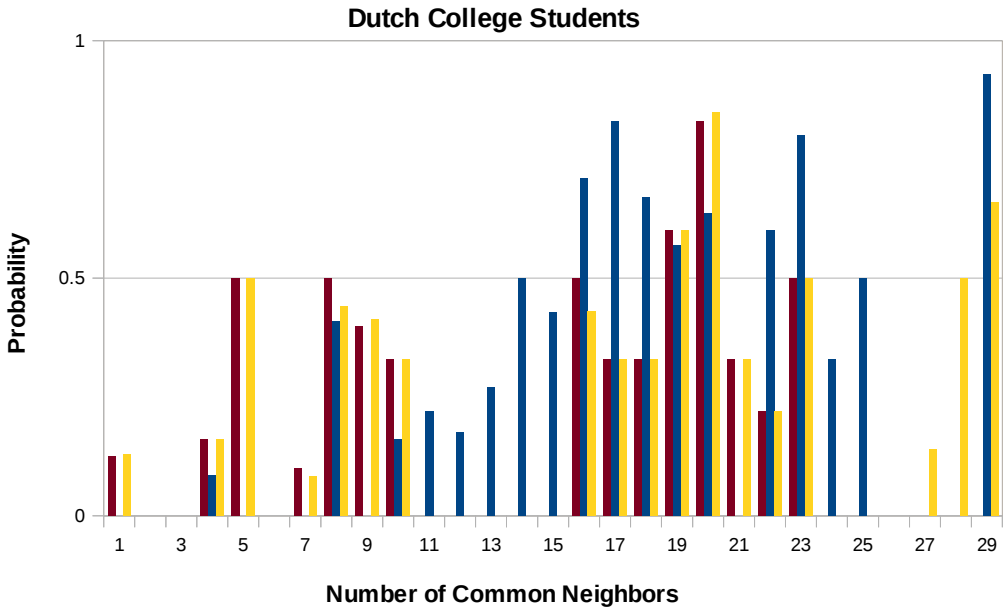



*Figure 9. Evaluation of the probability that two nodes will connect in the future, as a function of the number of common neighbors they have. The dark red bars represent the probabilities as they formed, when we used merely the friendship ties, while the yellow bars represent the configuration of the probabilities when all the given information, i.e. smoking habits and type of attending program, was included. The probabilities of the yellow bars, are product of the evaluation of all three snapshots and represent our estimation for the unknown possible future of the network.*

Figure 9 shows our evaluation of how likely it is for a new interaction among members to happen in the near future, based on the number of common neighbors they have. Given two snapshots of the network at time $t_1$ and $t_2$, we seek to accurately predict the edges that will be added to the network to a future time $t_3$. We can see the differences that appear, when we consume some (blue bar) vs all the available information (red bar) from the dataset. As we can see, in most cases, the extra information comes to fill in pieces and shift the probability towards a more certain outcome (values closer to 0 or 1).

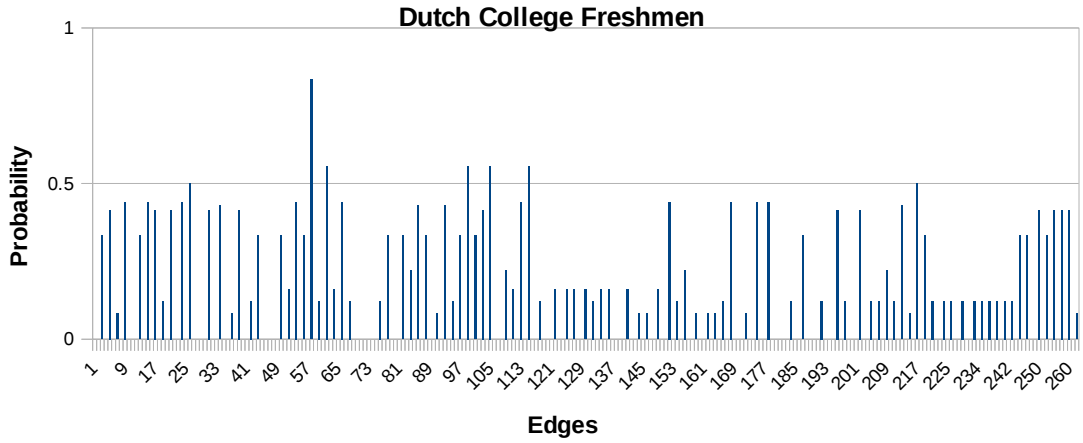



*Figure 10. Probability of an edge to appear and join a pair of nodes in the near future, based on our approach. This first predictor evaluated the probabilities, while having a limited image of the interactions existing in the network.*

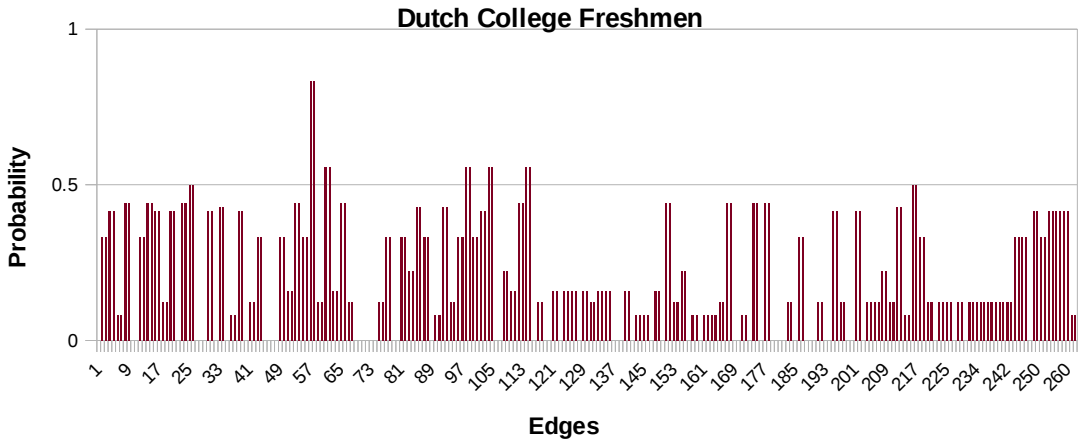



*Figure 11. Probability of an edge to appear and join a pair of nodes in the near future, based on our approach. The second predictor had a fuller image of the interactions existing inside the network.*

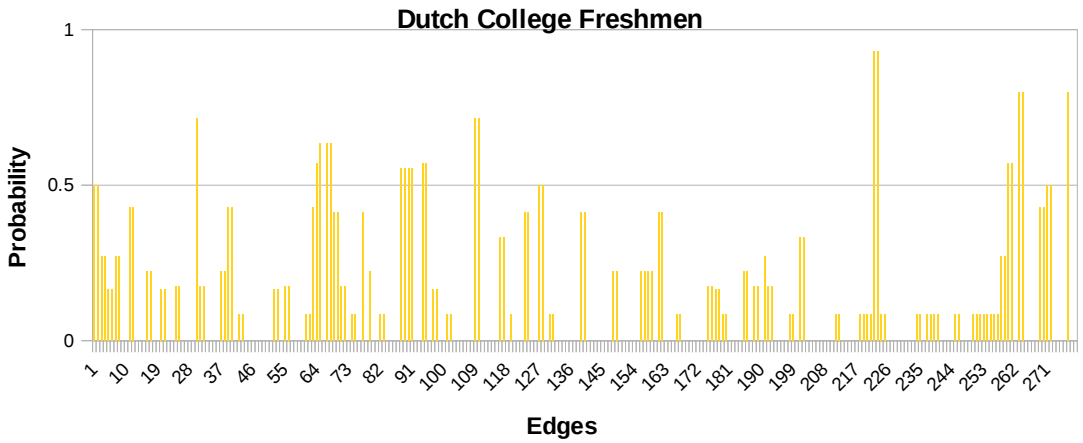



*Figure 12. Probability of an edge to appear and join a pair of nodes in the near future, based on our approach. The prediction for a currently missing edge, is based on all three snapshots of the network.*

Across Figure 10, Figure 11 and Figure 12, we can see the suggested outcomes for every single edge, based on our approach. In 87% of the cases, the first two predictors (Figure 10 & 11), which are products of the same two snapshots, are overlapping, showing similar behavior, while in the rest 13% they give reverse suggestions. We proceeded to compare the suggestions that the predictors created based on the two first snapshots of the network, to the new relationships that appeared in the third snapshot. As a matter of fact, the score (48%) of the predictor that was given the partly information was almost the same as the score (51,3%) of the predictor that had a fuller image of the interactions that exist in the network. The third predictor (Figure 12), that had as input all three snapshots of the network, gave the same predictions as the other two, in, averagely, 60% of the cases.

The set of the maximal frequent subgraphs that were obtained from the full image of the network, was, as expected, smaller, with more edges per set. On the other hand, the maximal frequent subgraphs of the limited image, where frequent subgraphs that appeared in the first set, but in this case they were not connected.

### 3.1.2 SG&R Law Firm

This dataset comes from a network study [*34*] of corporate law partnership that was carried out between 1988 and 1991 in New England, in a Northeastern US corporate law firm, referred to as SG&R. It includes the relations among the 71 partners and associates of the firm, covering friendly, strong co-working and advisor-advisee relationships.

For every attorney in the firm, it is also provided the number of years that they work with the firm, as well as some attributes, such as the individual's status in the firm, his/her gender, the location of the office he/she works in, his/her age, the law school they attended and their practice field. We took advantage of the given information about each attorney's years with the firm and we created 3 snapshots of the dataset.

Each node corresponds to an attorney and the edges represent a relationship between the attorneys. Each edge has also an attribute that indicates the type of the relationship. The average snapshot size, in terms of number of edges, is 818.6 . The largest snapshot contains 1,152 edges. The average number of edges connecting two nodes is 1.3 and the maximum number of edges between two nodes is 3. The average number of advisor-advisee relations that exist per snapshot are 190, while the average number of friendly and strong co-working relations are 242 and 386 respectively.

| | **Snapshot 1** | **Snapshot 2** | **Snapshot 3** |
|---|---|---|---|
| **Number of Triangles** | *9897* | *4659* | *797* |
| **Same-Type-Edges Triangles** | *2613* | *1013* | *482* |
| **Different-Type-Edges Triangles** | *7284* | *3646* | *315* |

*Table 4: SG&R Law firm. Number and types of triads.*

Table 4 shows the number of triangles that exist in every snapshot. The type of triangles that appear in the data, can either connect the nodes with edges of the same type (Figure 13), or the type of the edges will vary (Figure 14).

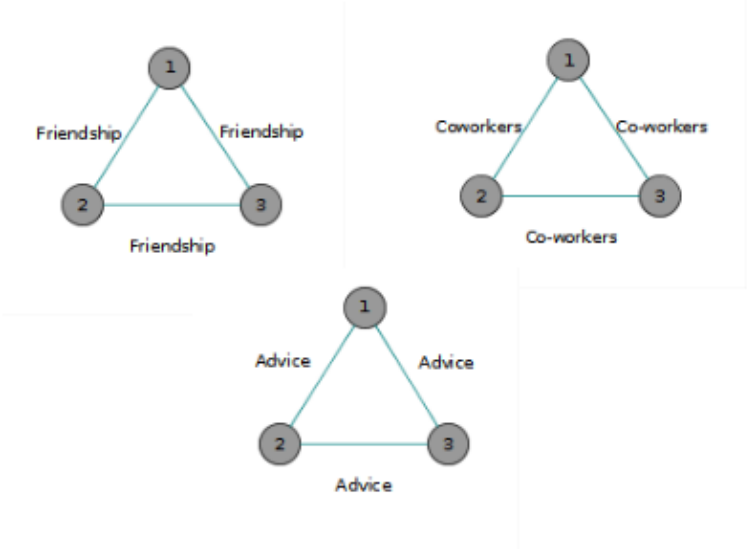



*Figure 13: SG&R Triangles. The possible configurations of triangle patterns between nodes that are interconnected with edges of the same type.*

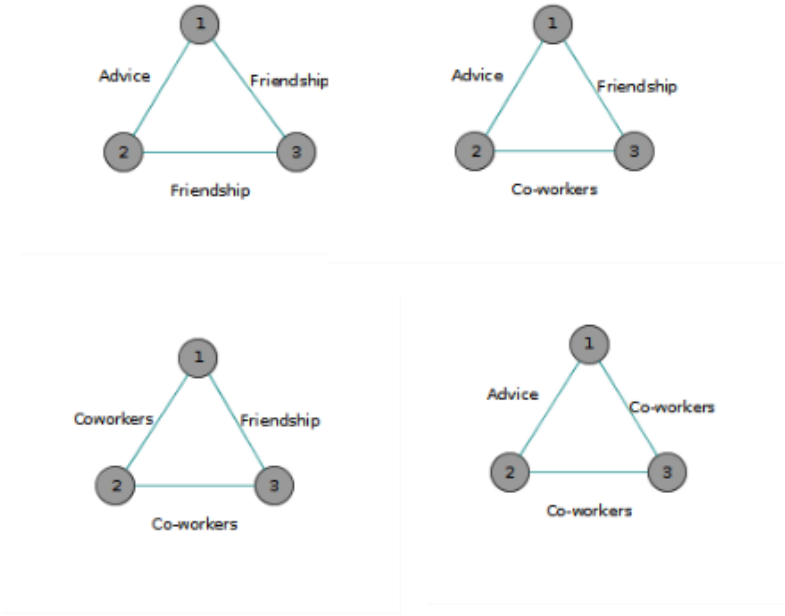



*Figure 14: SG&R Triangles. The possible configurations of triangle patterns between nodes that are interconnected with edges of different types.*

If we consider exclusively the advisor-advisee network, the triangle patterns that we extract are significantly fewer. In case we add only the ties that denote friendship, we get 2.3 times more patterns. While, assuming that we add only the ties that denote strong co-working relationship, we get 1.3 more patterns. Lastly, if we add them both and extract triangle patterns from the network that we described above, their number goes up to 5.9 times.

Within the advisor-advisee network, both friendships and strongly co-working ties are highly regular, with frequencies of 35% and 52% respectively. From these results we can conclude that attorneys have the tendency to seek for advice in their co-workers first and then in their friends. In a similar investigation within the network of strong co-working ties, the results indicate that strong co-workers have a relatively small tendency to maintain a friendly relationship as well, in a frequency of 12%.

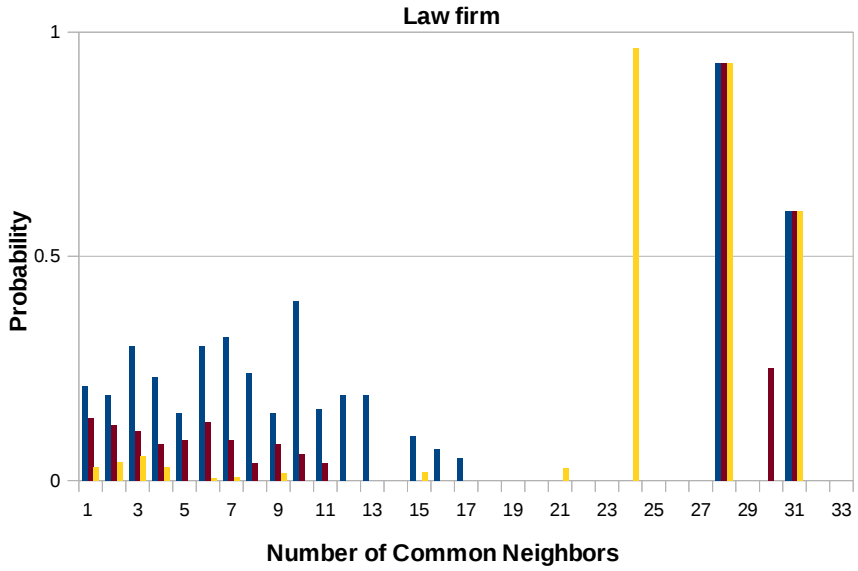



*Figure 15. Evaluation of the probability that two attorneys will connect in the future, as a function of the number of common neighbors they have. The dark red bar represents the probabilities as they formed, when we used merely the advisor-advisee ties, while the blue bar represents the configuration of the probabilities when all three types of ties were included. The yellow bar, represents the possibilities, as they formed when we used as input all three snapshots of the network.*

In Figure 15 we can see the differences between the estimated probabilities of the appearance of a relationship between two attorneys, based on the number of common people they have on their surroundings. The blue and the dark red bars are overlapping, with a few exceptions, meaning that the extra information that we discarded when we evaluated the dark red bars, is essential for the network. Actually, the predictions we made having the full image of the network, were 7% more accurate. The yellow bar represents the probabilities as they formed, when we utilized the edge and node information of all three snapshots of the network. In all three cases,

the highest probabilities of a relation to be formed, are between people with many neighbors in common.

Once again, the set of maximal frequent subgraphs acquired from the limited image of the network, where frequent subgraphs that we could find in the maximal frequent subgraph set of the full image.

### 3.1.3 Slashdot

Slashdot is a technology-related news website. In 2002, Slashdot introduced the Slashdot Zoo feature which allows users to tag each other as friends or foes. A positively signed link means that a user likes another user's comments, while a foe relationship means that a user finds another user's comments uninteresting. We used three snapshots of the network [*35*]. The first one is from November 2008, while the other two were obtained on February 16$^{th}$ and 21$^{st}$ , 2009. The size of the original dataset, in terms of number of edges is, averagely, 537,149 links per snapshot.

We investigated two parts of the original dataset. The size of the first and the second part, in terms of number of edges are, on average, 3,272 and 105,855 per snapshot, respectively.

| | **Snapshot 1** | **Snapshot 2** | **Snapshot 3** |
|---|---|---|---|
| **Number of Triangles** | *4404* | *4465* | *4248* |
| **Balanced Triangles** | *4155* | *4213* | *3997* |
| **Unbalanced Triangles** | *249* | *252* | *251* |
| **Type Of Triangle: + + +** | *3777* | *3810* | *3599* |
| **Type Of Triangle: + - -** | *378* | *403* | *398* |
| **Type Of Triangle: + + -** | *230* | *234* | *233* |
| **Type Of Triangle: - - -** | *19* | *18* | *18* |

*Table 5: Slashdot. Triangle pattern interconnections between 774 nodes.*

Table 5 shows the number of triangle patterns that exist in the first part of the dataset, along with their types. The count of positive and negative edges does not differ dramatically from snapshot to snapshot. The first one contains 2,913 positive and 353 negative links, the second one 2,679 positive and 71 negative links and lastly, the third one, 2,844 positive and 76 negative links. By allowing negative links in our dataset, we extract 1.24 times more triangle patterns. In all three snapshots, the three-mutual-friends type of triangles are overrepresented, while the three-mutual-enemies are extremely underrepresented. Also, in all three, the percentage of the balanced triangles is 94%, meaning that the conflicts' proportion in the network is stable.

Considering the configurations that balance theory suggests and whether a node tends to be balanced or not in the first two snapshots, we tried to predict, once more, the sign of the edges added in the third snapshot. The probability of the new edge to have the sign we predicted, was as high as 0.8 .

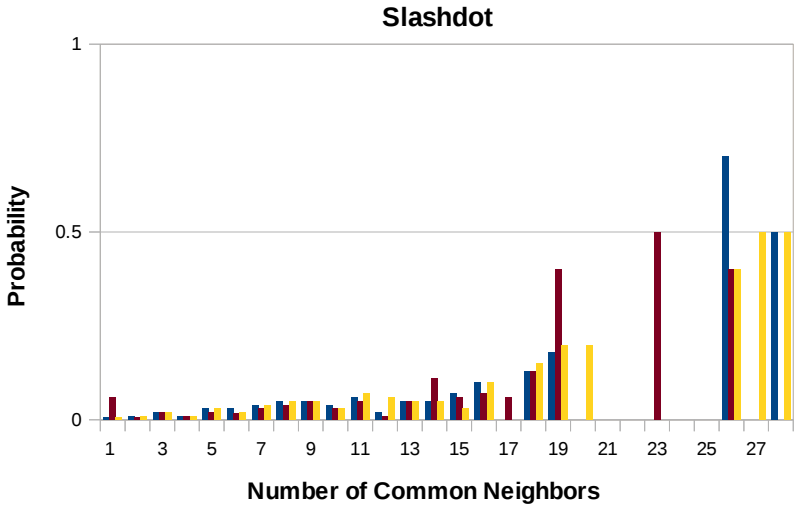



*Figure 16. Evaluation of the probability that two nodes will connect in the future, as a function of the number of common neighbors they have. The blue bars represent the probabilities as they formed, when we used both positive and negative links, while the dark red bars represent the configuration of the probabilities when we used merely the positive links. The yellow bar represents the probabilities that formed from the evaluation of the nodes and ties of all three snapshots.*

Figure 16 shows our evaluation of how likely it is for a new edge to appear in the near future, based on the number of common neighbors they have. Again, given two snapshots of the network at time $t_1$ and $t_2$, we tried to predict the edges that will be added to the network to a future time $t_3$. The predictors are highly overlapping, as in most cases they predict that an edge will not appear. We proceeded to compare the suggestions that the predictors created, based on the first two snapshots of the network, to the new relationships that appeared, or not appeared, in the third snapshot. Their scores were very close and they were higher than 90% right. It is useful to predict that an edge will not appear anytime soon in the network, but this approach is not useful for the prediction of the edges that will appear.

| | Snapshot 1 | Snapshot 2 | Snapshot 3 |
|---|---|---|---|
| **Number of Triangles** | *157818* | *157421* | *153853* |
| **Balanced Triangles** | *138219* | *137735* | *135321* |
| **Unbalanced Triangles** | *19599* | *19686* | *18534* |
| **Type Of Triangle: + + +** | *120851* | *120371* | *118204* |
| **Type Of Triangle: + - -** | *17268* | *17364* | *17117* |
| **Type Of Triangle: + + -** | *16698* | *16752* | *15730* |
| **Type Of Triangle: - - -** | *2901* | *2934* | *2804* |

*Table 6: Slashdot. Triangle pattern interconnections between 7736 nodes.*

Similarly, Table 6 shows the triangle counts and types of patterns that exist in the second part taken from the Slashdot dataset. The 78% of the edges of the first snapshot are positive, while 89% and 90% are positive in the second and the third snapshot, respectively. By allowing negative links in our dataset, we extract 1.3 times more triangle patterns. As before, the three-mutual-friends type of triangles are overrepresented, while the three-mutual-enemies are extremely underrepresented. In the first two snapshots, the proportion of balanced triangles against unbalanced ones is 87%, while in the third snapshot it slightly falls to 85%. That means that more conflicted relationships between nodes appear in the network. In this case, our attempt to predict the sign of the edges added to the network, was 3.5 times out of 5, successful.

Taking into account the above considerations, and applying them to both signed and positive only network, as before, the predictions for every particular edge were overlapping in several cases.

In both parts, the conclusion about the sets of maximal frequent subgraphs, was the same. The only-positive network's set, was a subset of the signed network's set.

# Chapter 4

## CONCLUSIONS AND DISCUSSION

In situations that we use graphs in order to model data in social networks and other applications, it is often important to allow the existence of multiple and/or signed edges, in order to have models that capture more than just a binary relationship.

In order to identify the importance of the type and the number of ties that compose a network in the process of pattern and link mining, we tested some mining techniques, in different kinds of social networks and we compared the results. The networks had different sizes and different types of edges. We investigated interdependencies between the nature of the different connections existing between two entities and how the lack of a connection affects the patterns mined from the network. The advantage of having signed and multi-attributed edges in a graph is that the type of patterns that we mine, are far more interesting, as they provide information about the type of the relationship that these entities have with each other, and give us a sense of the nature of each link and the behavior of each entity in the scope of a group. Moreover, the maximal frequent subgraphs that were mined, in terms of edges, were bigger, but the

set overall, was smaller, serving the benefits that the mining of maximal frequent subgraphs provides, which is to reduce the total number of mined subgraphs, to reduce the total mining time, to be able to reconstruct the non-maximal frequent subgraphs from them and encode the maximal structure commonalities within the graph [*36*]. Concerning the problem of link and sign prediction, the results were more hazy, as in some cases the existence of several types of relation, did not affect remarkably the results. On the other hand, we cannot ignore the increase in the cost and the complexity of the mine techniques, that the great rise in the number of edges causes.

In summary, this work presents how multiple and signed edges can affect the process of mining patterns and predicting behaviors in social networks. However, as we are dealing with an enormously growing field, with many applications and various techniques, the future of our work is naturally to test more techniques, along with larger and real world data, and compare the importance of the mined patterns against the increase of its cost, in order to get a better idea of the underlying possibilities and limitations.

# References

[1] Burke, M., Kraut, R.: Mopping up: Modeling Wikipedia promotion decisions. In: ACM Conference on Computer Supported Cooperative Work (CSCW), pp. 27–36. ACM Press, New York (2008)

[2] Brzozowski, M.J., Hogg, T., Szabá, G.: Friends and foes: ideological social networking. In: Human Factors in Computing Systems, CHI (2008)

[3] Evimaria Terzi, Marco Winkler: A Spectral Algorithm for Computing Social Balance. In: Algorithms and Models for the Web Graph, 8th International Workshop, WAW 2011 Atlanta, GA, USA, May 27-29, 2011 Proceedings

[4] Jure Leskovec, Daniel Huttenloher, Jon Kleinberg: Signed Networks in Social Media. In: *Proc. 28th ACM SIGCHI Conference on Human Factors in Computing Systems, 2010.*

[5] Fritz Heider. Attitudes and cognitive organization. Journal of Psychology, 21:107–112, 1946.

[6] M. Koyuturk, W. Szpankowski, and A. Grama, Biclustering gene-feature matrices for statistically significant dense patterns, *CSB'04*, 2004.

[7] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M. J. Zaki. Mining protein contact maps. BIOKDD, 2002.

[8] M. Koyuturk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. pages 200–207. ISMB, 2004

[9] Tibor Antal, Paul Krapivsky, and Sidney Redner. Social balance on networks: The dynamics of friendship and enmity. Physica D, 224(130), 2006.

[10] S. Wasserman, K. Faust. Social Network Analysis: Methods and Applications. Camb. U. Press, 1994.

[11] J. A. Davis. Clustering and structural balance in graphs. Human Relations, 20(2):181–187, 1967.

[12] http://www.adampalay.com/blog/2015/04/17/shakespeare-social-networks/

[13] M. Rocklin, A. Pinar: Latent Clustering on Graphs with Multiple Edge Types. In: Frieze, A., Horn, P., Pralat, P. (eds.) WAW 2011. LNCS, vol. 2777, pp. 38-49. Springer, Heidelberg (2011)

[14] Historical Developments and Theoretical Approaches in Sociology - Volume I

[15] Jiawei Han and Micheline Kamber: Data Mining: Concepts and Techniques (2006)

[16] Chuntao Jiang, Frans Coenen and Michele Zito: A Survey of Frequent Subgraph Mining Algorithms, The Knowledge Engineering Review, Vol. 00:0, 1–31. c 2004, Cambridge University Press

[18] Padgett, J.F. and C.K. Ansell, 1993. Robust Action and the Rise of the Medici, 1400-1434. American Journal of Sociology, 98, 1259-1398.

[19] Structural Models: An Introduction to the Theory of Directed Graphs [Frank Harary, Robert Z. Norman, Dorwin Cartwright]

[20] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. 2nd ed., Morgan Kaufmann, 2006

[21] http://moreno.ss.uci.edu/data.html,VAN DE BUNT--DUTCH COLLEGE FRESHMEN dataset

[22] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: Proceedings of ACM KDD, Las Vegas, NV, USA, August 2008

[23] J.-P. Eckmann and E. Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. PNAS, 99(9):5825–5829, April 2002.

[24] B. Kimelfeld and P.G. Kolaitis. The complexity of mining maximal frequent subgraphs. In: Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART. Symposium on Principles of Database Systems, PODS 2013, pages 13–24. ACM, 2013.

[25] Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *WWW'10,* pages 641-650, 2010.

[26] D. Cartwright and F. Harary. Structure balance: A generalization of Heider's theory. Psych. Rev., 63, 1956.

[27] R. S. Burt. The network structure of social capital. Research in Organizational Studies, 22:345–423, 2000.

[28] Ronald L. Breiger. The duality of persons and groups. Social Forces, 1974.

[29] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In: ICDM'03, 2003.

[30] X. Yan and J. Han. Closegraph: Mining closed frequent graph patterns. KDD'03, 2003.

[31] Wasserman, S., and Faust, K. (1994). Social Network Analysis: Methods and Applications.

[32] Social Network Analysis: A Handbook by John Scott, London: Sage (2000).

[33] Van de Bunt, Gerhard G., Marijtje A. J. Van Duijn, and Tom A. B. Snijders. 1999. Friendship Networks Through Time: An Actor-Oriented Statistical Network Model. In: Computational and Mathematical Organization Theory 5: 167-192 .

[34] Emmanuel Lazega, The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press (2001).

[35] J. Leskovec, D. Huttenlocher, J. Kleinberg: Signed Networks in Social Media. 28th ACM Conference on Human Factors in Computing Systems (CHI), 2010.

[36] Jun Huan, Wei Wang, Jan Prins, Jiong Yang, Spin: Mining maximal frequent subgraphs from graph databases. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pages 581–586, New York, NY, USA, 2004.