

**EXTENDED BALANCE THEORY
AND CONVERGENCE MODEL
IN SOCIAL NETWORKS**

By

Yi Qian

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
Major Subject: COMPUTER SCIENCE

Examining Committee:

Sibel Adalı, Thesis Adviser

Elliot Enshelevich, Member

Petros Drineas, Member

Rensselaer Polytechnic Institute
Troy, New York

March 2013
(For Graduation May 2013)

© Copyright 2013
by
Yi Qian
All Rights Reserved

CONTENTS

LIST OF TABLES	iv
LIST OF FIGURES	v
ACKNOWLEDGMENT	vi
ABSTRACT	vii
1. INTRODUCTION	1
2. BACKGROUND AND RELATED WORK	6
2.1 Strong and Weak Structural Balance Theory	6
2.2 Social Network Convergence	8
2.3 Edge Sign Prediction	9
3. EXTENDED BALANCE THEORY	12
3.1 Tolerance and Two Principles	13
3.2 Measuring Relationships by Total Ordering	15
3.3 Relation Distance and General Expression of Balance	19
4. CONVERGENCE MODEL	22
4.1 Convergence Model	22
4.2 Predictability of Convergence Model	24
5. ALGORITHMS FOR STRESS MAJORIZATION	34
5.1 Review of stress majorization	34
5.2 Stress Majorization in Large Social Networks	35
6. EXPERIMENTS AND RESULTS	41
6.1 Data and Algorithm	42
6.2 Results	43
6.3 General Edge Prediction	44
7. CONCLUSIONS AND FUTURE WORK	47
REFERENCES	48
APPENDICES	
A.	51

LIST OF TABLES

3.1	Tolerance rules in structural balance theory.	17
3.2	Five types of relations and their interpretations, displayed in the ordering, i.e. $s+ \preceq w+ \preceq O \preceq w- \preceq s-$	17
3.3	Tolerance rules with strong, weak and neutral relations.	18
3.4	Imbalanced triadic structures in the presence of strong and weak ties, and arguments for the stress in the relation.	19
3.5	Constraints on the boundary parameters.	21

LIST OF FIGURES

1.1	Four different cases of triangular relationship.	1
3.1	Partitioning of distance domain by boundary parameters $\{b_{s+}, b_{w+}, b_{w-}, b_{s-}\}$: distances within $(0, b_{s+}]$ are strong positive; (b_{s+}, b_{w+}) weak positive, $[b_{w+}, b_{w-}]$ neutral, (b_{w-}, b_{s-}) weak negative, and $[b_{s-}, \infty)$ strong negative.	20
4.1	The demonstration experiments of Propositions 1, 2 and 3. Each blue ball has a white number i in it, indicating node i . The initial state of each triad is shown in figures on the left, and the corresponding converged layouts on the right.	26
4.2	The demonstration experiments of Corollary 1 and 2. Each blue ball has a white number i in it, indicating node i 's location in its converged layout.	30
4.3	The demonstration experiments of Proposition of Aggregation. Each blue ball has a white number i in it, indicating node i . The initial state of each triad is shown in figures on the left, the corresponding converged layouts in 2-D space is shown in the middle, and the corresponding converged layouts in 3-D space is shown on the right.	31
6.1	The ROC curves are drawn upon distances of hidden edges, generated by SM and FD. (1) (2) (3) correspond to the ROC curves generated from Epinions, Slashdot and Wikipedia datasets respectively.	44
6.2	The histograms are drawn upon distances of neutral testing edges, generated by SM and FD. (1) (2) (3) correspond to the histograms generated from Epinions, Slashdot and Wikipedia datasets respectively.	45

ACKNOWLEDGMENT

It has been a great privilege to spend two and a half years in the Department of Computer Science at Rensselaer Polytechnic Institute. One of the joys of a completion is to look over the journey past and remember all the people who have helped and supported me along this long but fulfilling road.

My first debt of gratitude must go to my advisor, Dr. Sibel Adalı. She patiently provided the vision, encouragement and advise necessary for me to proceed the research journey. I hope that I could be as lively, enthusiastic, and energetic as Sibel and to someday be able to command an audience as well as she can. Not only that she has always been an inspirational and experienced advisor to me throughout my graduate school years, but also she supports my choice in exploring various subjects with great freedom. I could not have asked for a better role model as a researcher, as well as a great person.

I would like to thank my committee, Dr. Elliot Anshelevich and Dr. Petros Drineas for their support and guidance. Special thanks to Dr. Elliot Anshelevich to his many valuable suggestions. The guidance has served me well and I owe them my heartfelt appreciation.

I wish to thank my parents, Guahua Qian and Pinhong Jiang. Their love provided my inspiration and was my driving force through the years. My girl, Feifei Bian, whose love and encouragement brings me all the way here. In our four-year relationship, we bear the brunt of frustration and share the joy of success. Without the love and support of them, I would certainly not be able to finish this thesis.

I also would like to thank T. DuBois for sharing the implementation details of his algorithm.

ABSTRACT

Structural balance theory studies the influence between positive and negative relationships in a social network. Researchers regard structural balance as a sound and universal concept in social networks. We first observe the fact that current structural balance theory does not distinguish between varying strengths of relationships, which is shown to be of good significance in many social interactions. We address this problem by analyzing the social and psychological source of imbalance, and establish an extended theory that defines balance in networks with varying relationship strengths. We show how balance is reasoned when strengths of relationships are measured by either a totally ordered set, or by numerical values. Our extended balance theory is shown to be a generalization of the current structural balance theory.

The convergence model studies how an imbalanced network evolves towards new balance. The assumption behind our model of convergence is that in resolving tensions within imbalanced relationships, people tend to avoid the effort involved in changing relationships if possible. The introduction of extended balance theory allows us to formulate the convergence problem of a social network as a Metric Multidimensional Scaling (MDS) optimization problem.

Modeling the dynamic nature of social networks, the convergence model inherits a predictive power on unknown relationships. We show that a list of famous social network phenomena, such as “If two people have more common friends, then they are more likely to become friends”, hold under the convergence model. Moreover, we use the convergence model to study the *edge sign prediction problem*. In a social network with positive and negative signed edges, the task of sign prediction is to infer the signs of a small fraction of “hidden” edges based on the information from the rest of the network. We show why and how the convergence model can be applied in prediction tasks over real online social network datasets. Stress majorization technique is used to compute the convergence, and our method consistently matches and outperforms the state of art.

CHAPTER 1

INTRODUCTION

In a social network with signed edges, positive edges represent friendship while negative edges represent antagonism. An important problem in the study of social networks is to understand the tension between these two forces [1]. Structural balance is a concept studying the influence of signed relationships on each other. The theory of structural balance has its origins in the work of Heider [2]. Cartwright and Harary then formalized Heider's idea and formulated the overall structure of a network with signed edges [3]. Davis [4] further gives a generalized theory by a relaxed balance assumption. In essence, structural balance theory considers certain configurations of positive and negative edges as socially and psychologically more plausible than others. In balance theory, a triangle, i.e. triad, is the smallest structural unit. It consists of three pairwise relationships between three people, and there are four distinct ways to label each relationship as positive or negative; see Figure 1.1

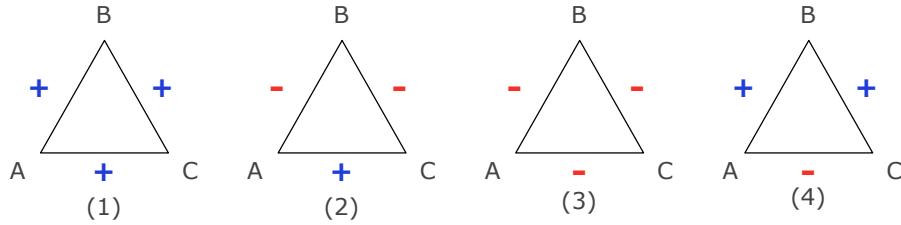


Figure 1.1: Four different cases of triangular relationship.

It is argued that triads (1) (2) are more socially and psychologically reasonable than than triads (3) and (4) [2], [3]. The configurations of positive and negative relations in triads (3) (4) will bring tension that makes such triadic relationship unstable. Davis points out that the stress inside triad (4) is much more significant than it is in triad (3), and hence proposes a generalized balance theorem [4].

The structural balance theory provides a concise and elegant description of influence between interpersonal relationships. Moreover, it illustrates a nice connection between local configuration and global structural property in social networks.

The key result of Cartwright-Harary’s work is the structure theorem [3]: if a network is balanced, the nodes can be partitioned into two subsets such that all positive relationships are within subsets and all negative relationships are between subsets. Davis draws a similar theorem in the case when triad (3) in Figure 1.1 is relaxed as balanced [4].

In recent research in online social networks, network balance plays a key role in many applications. One specific problem of interest is inferring new relationships and make recommendations based on existing relationships. We observe that the concept of structural balance is widely applied when developing algorithms for this type of prediction task. For example, Leskovec et al. generate a class of triad features in their prediction algorithm, i.e. a pair of relationships constraining a third relationship [5]. They also use the structural balance as a touchstone to see the congruence between their practical results and the long studied theory. Another algorithm [6] makes use of similar assumptions without explicit mention of balance theory in which trust and distrust relationships are mapped to metric distances on a continuous range. To some degree, the strong prediction performance of these algorithms justifies the structural balance theory. However, we will see the theory has its own weakness.

The arguments behind the balance theory imply a type of tendency towards balance for general social networks. That is, every social network will evolve to a state where every triad in it is balanced. The assumption by itself is congruent and intuitive; unstable triadic relationships (triads (3) (4) in Figure 1.1) will break and change to stable ones because of the inside tension. Regardless of other potential factors, a social network will inevitably reach a balanced state. As the “balance process” is by nature a convergent process, every social network is expected to stay at a largely structurally balanced state. However, for most observed signed structures for social groups, exact structural balance does not hold [7].

Despite the fact the structural balance theory is insufficient of empirical support, we regard the balance tendency a sound and natural description of social networks. In the physical world, an object stays at a state of minimum energy. Analogically, in social networks, does a triadic relationship stays at a state with

minimum inner tension? We regard the structural balance theory as an incomplete description of balance in social networks. The current theory only deals with binary relationships while relationships in practice vary in degrees. While it is shown that strength is an important factor in many different social phenomena [8], the structural balance theory does not distinguish between relationships of the same sign but different strengths. In fact, when Cartwright and Harary first formalized the theory of structural balance, they also had the following suggestions [3]:

“...Obviously, however, many relationships of interest to psychologists (like liking, for example) exist in varying degrees. The fact means that our present use of graph theory can treat only the structural, and not the numerical, aspects of relations. While our treatment is thereby an incomplete representation of the strength of relations, we believe that conceptualization of the structural properties of relations is a necessary first step toward a more adequate treatment of the more complex situation.”

We take steps to analyze the social and psychological source of imbalance, and propose two fundamental principles regarding triadic balance in the general sense. Based on the principles, we establish an extended balance theory that deals with relationships with varying strengths. Heider’s structural balance theory will be shown to be a special case under the generalized theory.

In the second part of the thesis, we study the problem of social network convergence. The balance theory defines what is a balanced state of a social network, but does not describe the dynamic aspect of the “balance process”. The problems of what each imbalanced triad will change to over time, and globally how an imbalanced network would evolve towards balance has not been studied in literature. It is plausible to have a decisive argument on whether triad (4) in Figure 1.1 will change to triad (1) or triad (2), given the imperfect empirical evidence of structural balance itself. That is, if Bob’s two friends Alice and Chris cannot get along with each other, whether Alice and Chris will resolve conflicts and become friends, or Bob teams up with one of them? In real life, questions like this really depend on how close Bob and Alice (Chris) is and how much conflict there is between Alice and Chris. For example, if the two friendships are very close while the conflict is subtle,

we expect to see Alice and Chris become friends. On the contrary, if Bob is only close to Alice while the hatred between Alice and Chris is intense, it is likely that Bob will team up with Alice. With the extended balance theory, the problem becomes non-trivial and potentially solvable. We call such dynamic “balance process” the **social network convergence**.

The convergence model helps complete the “process” of network balance, as it models the principles of inter-relationship influence aspect of the network evolution. Due to its dynamic nature, the convergence model should inherit a predictive power over new relationships. For example, the strong triadic closure states that two people with a common close friend are very likely to friend each other in the future [8]. We propose some famous predictive properties in social networks, such as “two people with more common friends are more likely to become friends” as touchstones of the model, and show they hold under the convergence model. Both empirical experimental results and proofs are given.

To further justify our theory empirically, we use the convergence model to study the *edge sign prediction* problem. The *edge sign prediction* problem uses information of existing relationships to infer hidden relationships. To this date, many algorithms for this prediction problem are based on machine learning methods. These methods generate multiple structural features from the known relationships, and then use these features to classify the unknown relationships [5], [6], [9]. We show why and how the convergence model can be applied in the prediction tasks over real online social network datasets. Our method consistently matches and outperforms the state of the art.

To summarize, we make the following unique contributions in this thesis:

- We first introduce a new extended balance theory that allows arbitrary relationship strengths. We express balance with two simple principles that preserve the unique meaning of a positive, and a negative relationship. We show how balance can be reasoned when the strength of relationships are expressed as either discrete categorical values, as pairwise comparisons or as metric distances using the same two principles. Our method is novel, has sound social and psychological basis and captures the classical balance theory as a special

case.

- We propose a convergence model, describing how an imbalanced network evolves towards new balance. The assumption behind our model is that in resolving tensions within imbalanced relationships, people tend to avoid the effort of changing relationships if possible. The introduction of extended balance theory allows us to formulate the convergence problem of a social network as a Metric Multidimensional Scaling (MDS) optimization problem.
- We show how the convergence model can be used to predict edge signs in social networks, and justify our theory through experiments on real datasets. Stress majorization is applied to solve MDS [10], and our method consistently matches and outperforms the state of the art. In addition, we show promising results towards providing solutions for the harder *link prediction problem*.

Following the introduction, the thesis is organized as follows. Chapter 2 concisely reviews the structural balance theory and related work on *edge sign prediction* problem. Chapter 3 generalizes structural balance theory to allow relationships to have varying strengths, and establishes an extended balance theory to capture the general concept of balance. Chapter 4 discusses the social network convergence, and models it as an MDS problem. Some well-known social network phenomena are illustrated and proved under the convergence framework. Chapter 5 reviews the stress majorization technique to solve the MDS problem. Due to its high computational cost, two algorithms are introduced so that convergence model can be applied to large social networks. Chapter 6 examines the performance of the convergence model on the *edge sign prediction* problem. Chapter 7 concludes the thesis.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Strong and Weak Structural Balance Theory

The concept of structural balance is based on theories in social psychology dating back to the work of Heider [2], and generalized and formulated by Cartwright and Harary [3],[11],[12]. When modeling relationships between pairs of individuals, positive relationships are representative of liking, loving, valuing or approving someone, and negative relationships are representative of disvaluing, disapproving or negatively valuing someone [3]. Structural balance theory considers social networks with binary relationships, and argues that certain configurations of a triadic relationship are socially and psychologically more plausible than others. see Figure 1.1.

- Given three people Alice, Bob and Chris, it is natural for them to be mutual friends of each other (triad (1) in Figure 1.1).
- Similarly, a relation in which two friends, Alice and Bob have a common enemy Chris, is also natural (triad (2) in Figure 1.1).
- The other two configurations of triangles introduce psychological stress or tension into relationships. It is “stressful” for Alice and Bob, and Bob and Chris to be friends, but Alice and Chris to be enemies (triad (4) in Figure 1.1). Such stress is a kind of implicit force that will either push Alice and Chris to become friends, or else force Bob to side with one of Alice and Chris [1].
- Similarly, there is stress that motivates two of the three people to “team up” against the third one in the situation when Alice, Bob and Chris are mutual enemies against each other (triad (3) in Figure 1.1) [1].

In [3], triads (1) (2) in Figure 1.1 are referred as balanced while triads (3) (4) in Figure 1.1 are referred as imbalanced, which is referred as the Strong Balance Theory (SBT) in [1]. It is argued people tend to reduce the psychological dissonance resulting from imbalanced triads. Hence, balance theorists argue that in real social

networks, imbalanced triads are unstable and subject to changes towards balanced structure.

In a complete network, all pairs of people are connected to each other by positive or negative links. We call it a balanced network if all triads in it are of balanced structures, i.e., all triads are of the form (1), or (2) from Figure 1.1. Cartwright and Harary also discuss the global structure of a balanced complete network [3].

Theorem 1 *If a labeled complete graph is balanced, then either all pairs of nodes are friends, or else the nodes can be divided into two groups, X and Y , such that every pair of nodes in X like each other, every pair of nodes in Y like each other, and everyone in X is the enemy of everyone in Y .*

Examples of SBT and Theorem 1 has been illustrated with some anecdotal examples from international relations. For example, the United States surprisingly supported Pakistan in 1972. At that time, U.S. was trying to improve relations with China while China and Pakistan was close because of their common enemy India and USSR. We see a relatively clear dichotomy of these countries.

James Davis argues that the latent stress of the two imbalanced triads (triads (3) (4) in Figure 1.1) in Cartwright-Harary's theory is fundamentally different [4]. In (4), we have the **stress source** of a person whose two friends do not get along; in (3), there is the **possibility** that two of the nodes will ally themselves against the third. Davis points out that in many settings, the imbalanced factors in (4) may be significantly stronger than it is in (3) [4]. More often, we see two people with a common friend try to reconcile their differences (triad (4) in Figure 1.1), while there is weaker motivation that leads two of three mutual enemies to become ally. The balance theory that only considers triad (4) in Figure 1.1 as imbalanced is referred to the Weak Structural Balance Theory (WSBT).

Just like SBT, WSBT enjoys a similar global structural property. Namely, if a complete network is weakly balanced, then it can be partitioned into multiple groups such that nodes within the same group are mutual friends, and nodes belonging to different groups are enemies [4].

2.2 Social Network Convergence

The concept of “evolution of social networks” is not new. Researchers have used longitudinal data to model such evolution. Leinhardt [13], Wasserman [14], Leenders [15] and Snijders [16] have proposed to use continuous-time Markov chains as a model for evolution of social networks. Particularly in [16], the network evolution is modeled as the consequence of the actors making new choices, or withdrawing existing choices, on the basis of functions of utility, with fixed and random components, that the actors try to maximize. The change in the network is modeled as the stochastic result of the network effects.

Doreian suggests that social networks change through the operation of coherent social processes [7]. One coherent theory of such “evolution process” is the balance theory, as imbalanced structures in a network are unstable and subject to change. However, as Doreian points out, the “network evolution” is influenced both by the balance process and individual characteristics [7]. On the one side, balance process regulates the mutual influence over relationships at a structural level; certain structures (triads) will be forced to change due to the latent social and psychological stress. On the other side, relationships are changed following the interests or characteristics of the individuals in a relationship. For example, one may establish positive relationship with someone who provide valuable information, or build negative relationship with someone who have conflicted interests. Take both evolution factors into account, the evolution of a social network qualifies as an autonomous system in which balance and individual interests coexist. Due to the unpredictability and complexity of individual acts, it is unlikely to model the evolution process by a concise and unified theory. Hence, we only discuss the balance aspect of the network evolution in this thesis. Since the stress within imbalanced structures always pushes the network towards balanced states, every social network has a tendency towards balance. Regardless of other factors, the balance evolution of a social network would converge at one point when the balance is reached. We therefore call the balance aspect of the evolution as **social network convergence**.

The convergence of social network has not drawn much attention under the framework of binary relationships. With the extended balance theory, however, the

problem becomes interesting and rather complicated, as changes in relationship can be numerical.

2.3 Edge Sign Prediction

The *edge sign prediction* problem is studied as an application, as well as a justification, of SBT theory. Namely, suppose we are given a social network with positive or negative edges, but a small fraction of the edge signs are “hidden”. How can we predict the signs of hidden edges with the information provided by the rest of network?

Guha et. al. [9] introduce one of the earliest methods that addresses the propagation of both trust and distrust. To solve this problem, they propose the concepts called direct propagation, co-citation and backwards propagation, and compute trust propagation by repeating matrix operations that combine the three types of propagations. They report an overall 85% prediction accuracy over data samples from Epinions that has equal number of positive and negative edges.

Leskovec, Huttenlocher and Kleinberg [5] first formulate the *edge sign prediction* problem. They conduct a series of experiments on three large datasets: Epinion, Slashdot and Wikipedia based on a machine learning framework. In particular, they collect two classes of features, one of which is based on degree and the other is based on triads. These relatively local features form a high dimensional space on which they perform standard machine learning methods and perform edge sign predictions. Closely related to our work, they also interpret some of their results in terms of Cartwright-Harary’s balance theory [3], but unlike our work, they do not use balance theory as a starting point of their approach.

The recent work by DuBois et. al. [6] is also related to this thesis from an algorithmic point of view. This work stands out as it provides very good prediction performance for the edge sign prediction problem: 80 – 90% accuracy on all of the three datasets used in [5] for both positive and negative edges. In this paper, the authors map trust and distrust relationships to metric distances: the larger the distance is, the more negative (less positive) the relation is. The proposed method computes two features for each signed edge: the first one is based on path probability

(PP, $O(n^2)$) [17] and the second one uses a force directed algorithm (FD, $O(kn)$ at each iteration where k is the average degree of the network) [6].

In this thesis, we show that some of the assumptions underlying this algorithm can be formally defined as part of a general structural balance theory that not only works for simple positive and negative edges, but also takes into account relation strength when applicable. Being able to deal with strength also enables us to state the explicit optimization criteria in the metric space for the graph drawing problem. As a result, we are able to compare the prediction performance with respect to an optimal placement of nodes according to our theory.

Notice that both the force directed algorithm (FD) and stress majorization (SM) that we use in this thesis have been used in the field of graph drawing [10]. In FD, an attractive force is assigned between endpoints of each positive edge and a repelling force is assigned between endpoints of each negative edge. Nodes are initially randomly laid out, and the system is simulated until a stable equilibrium is reached when the total kinetic energy is below certain threshold. The relation between every pair of nodes is represented by the distance between the two end nodes in the stable layout of the network.

While FD is simple to implement, it operates on a local pairwise level, instead of a global level. This can lead to problems if the local forces end up not being sufficient to hold small groups together. Alternatively, if negative forces are too high, then the network may continuously expand in space and the algorithm may never converge. As a result, such a method requires carefully tuned parameters for a specific network to work well. In contrast, SM is a mature approach that guarantees monotonic convergence for drawing graphs. Moreover, in [6], there is no force between pairs of unconnected nodes which can result in unintuitive distances for such pairs. In fact as we show in our results, the FD method maps unconnected nodes to a predominantly positive range. This presents a problem for using this algorithm for solving the harder *link prediction problem* [18], which predicts the presence of a positive relationship between two arbitrary nodes. Link prediction is a harder problem since networks are often sparse and one needs to find the few edges that are true positive or negative links with high probability. We show that our

results are very promising on this front.

In addition, to the best of our knowledge, none the existing methods provide a way to study the principles underlying positive and negative relationships in very large networks with varying degrees of relationship strengths. It is unclear to which degree SBT or WSBT balance theory is valid for many large networks in which some or most relationships are simple acquaintances [8] instead close friendship relationships. An acquaintance may not result in the same type of structural constraints. For example, if Alice knows Bob, and Bob knows Chris, but Charlie dislikes Alice, this may not cause much stress in the existing relationships if Alice, Bob and Chris rarely spend time with each other, i.e. their relationship is not strong. However, there are still some implications for the network overall when we consider acquaintances as well as friendships. We examine those in the next chapters and provide a flexible theory of balance that generalizes WSBT. We show that our theory allows us to formulate convergence as an optimization problem, which can be solved by stress majorization and illustrate that our algorithm achieves better performance than those cited in the literature [6] while also providing a principled way to approach the *edge sign prediction* problem.

CHAPTER 3

EXTENDED BALANCE THEORY

In Heider’s balance theory [2], relations are restricted to binary values (+/−). When Cartwright and Harrary first formalized the theory of structural balance, they also suggested that relationships of interest exist in varying degrees, and that their theory is built on the incomplete representation of strengths of relations [3]. Tie strength is a well-studied concept in social psychology. A person may have close friends as well as acquaintances, strong and weak ties. A strong tie may represent a deep trust relation spanning many constructs in high risk situations, while a weak tie may be trusted mainly for low risk situations or for specific constructs like providing private information [8].

To model this distinction, we consider a scenario where relationships have varying strengths. For example, a strong positive link represents a close friendship or family tie, and a strong negative link represents hatred. The new balance theory in this general scenario will be called extended structural balance theory, or ESBT for short.

Balance theory deals with the influence between interpersonal relationships, and hence a triad is the smallest unit. In a complete network in which every pair of nodes has a mutual relationship, the multilateral relationship of a subset of participants can be captured as a composition of all triads involved. When a social network is incomplete, by letting the “no link” be a type of relationship, the same argument still holds. In other words, triad not only defines the smallest unit structure of social networks, but is also able to characterize every social structure by the composition of all triads involved. As an example, the *balance theorem* in SBT (WSBT) has shown the power of triad in capturing the global structure of a network. No matter how the relationship is measured (binary or varying in strengths), the center of discussion of a balance theory lies in triads: a social network is balanced if every triad involved is balanced.

We consider a type of **neutral relationship** as one that is unbiased, which will

be denoted as O . Basically, a neutral relationship is a non-negative and non-positive relationship, corresponding to no opinion and no bias. As a result, “no link”, or “no relationship”, is a type of neutral relationship. Clearly, every combination of three nodes in a social network forms a triad if neutral relationship is considered.

Let the collection of relations with strengths be E . With the introduction of neutral relations, E can be partitioned into three subsets: positive relationships P , negative relations N and neutral relations O . An edge (A, B) and a relation with associated strength e will be used interchangeably. We call e_1 is more positive (less negative) than e_2 if

1. e_1 is positive and e_2 is negative;
2. e_1, e_2 are positive, and e_1 is stronger than e_2 in strength;
3. e_1, e_2 are negative, and e_1 is weaker than e_2 in strength.

A triad is usually denoted as (A, B, C) .

3.1 Tolerance and Two Principles

Reviewing the arguments in Heider’s balance theory [2], one concludes that the latent stress, or tension, causes a triad to become imbalanced. For example, in triad (4) of Figure 1.1, it is stressful for Alice and Chris to stay antagonistic to each other with Bob as a common friend, while Bob also feels stressed by staying friends with both Alice and Chris. As it is argued by Davis [4], the situation in triad (3) of Figure 1.1 is fundamentally different. In (3), there is a **possibility** that two of the nodes will ally themselves against the third. In a word, the underlying stress in imbalanced triads is the driving force that propels relationship changes. The next questions is when and how such stress rises.

We further our discussion by describing stress in terms the influence of relationships on each other within a triad. In a triad, two of its relations cause influence over the third one. Such influence restricts the range of the comfortable relations the third relation may have; if the relation goes out of the range, tension occurs and participants will suffer from stress. Participants will seek relationship changes

to resolve stress. We call such range of relations **tolerance**. Using the notion of tolerance, we can interpret network balance at a finer level.

For each relationship between a pair (A, C) of nodes, its tolerance is constrained by the triads (A, C) is part of. If consider a single triad (A, B, C) , (A, C) 's tolerance is determined by the strengths of relations (A, B) and (B, C) . We state the following two fundamental principles regarding tolerance:

Principle 1 (Transitivity of positive relationships.) *Let (A, B, C) be a triad of interest, and (A, B) , (B, C) be positive. If (A, B) and (B, C) are more positive, (A, C) 's tolerance will be limited to more positive relationships.*

In other words, the fact that B are friends with both A and C provides the freedom for A and C to become friends; and there is stress on A and C to get close. The stronger the relation between (A, B) and (B, C) , the higher the chance between (A, C) to be connected (more) positively, and the resulting tolerance is restricted to be more positive.

The stress that is based on positive relations has been frequently defined by SBT and WSBT. Positive relations in a triad cause stress for the remaining relations to be positive. As a result, both in SBT and WSBT, a balanced network consists of communities that are connected to each other with positive ties. When we consider the strength of relations, we generalize this by saying that the more positive two of the relations are in a tie, there is less tolerance for non-positive relations.

Furthermore, there is a point when the strengths of the two positive relations (A, B) and (B, C) are strong enough such that it will be imbalanced for (A, C) to remain unfriended, i.e. neutral. This observation is inspired by the “strong triadic closure” in [8]. In trust literature, for example, this is often referred as the “transitivity of trust” though transitivity is also used in other contexts.

Principle 2 (Heterophily in relationships.) *Let (A, B, C) be a triad of interest. If the difference between (A, B) and (B, C) is larger, (A, C) 's tolerance will be limited to more negative relationships.*

Given individuals A, B, C in a network, if the relationship between (A, B) and the relation between (B, C) differs to some extent, then the tolerance is geared towards

the negative. Furthermore, the more different the strength of the relationships are, the tolerance is geared towards more negative relationships.

The second type of stress is an interpretation of homophily. We note that homophily, i.e. having common friends or enemies, may sometimes cause stress (in $+, +, +$) but sometimes it does not (in $-, -, -$ for WSBT). However, lack of homophily, which we call heterophily does cause stress. For example, consider the case $+, +, -$ for (A, B) , (B, C) and (A, C) . There is stress on (A, C) to be positive due to transitivity. But, there is also stress on (A, B) or (B, C) to be negative. Either way, the result will be more desirable: either all being friends, or having two friends with a common enemy. We call the second type of stress the principle of heterophily. The more different the ties are, the more pressure there is for the tie to be negative. At a point when the difference between (A, B) and (B, C) is significant enough, we argue that a positive relationship for (A, C) will cause imbalance. This is inspired by the observation that two people who have severely conflicted relationships with a common neighbor, e.g., one is the other's close friend's enemy, are not likely to friend each other.

To summarize, the tolerance of each edge in a triad is determined by the other two relationships. A triad is balanced if all three relationships in it are in the range of the corresponding tolerance. The concept of tolerance and its two principles help interpret balance in terms of relations with strengths precisely.

Definition 1 (Balance) *A triad A, B, C is balanced if for all pairs (A, B) , given the tolerance $T(A, B)$ of (A, B) with respect to $(B, C), (A, C)$, we have that (A, B) is in $T(A, B)$.*

Given a network G of relationships, G is said to be balanced if for all triads in the network are balanced.

3.2 Measuring Relationships by Total Ordering

To give concrete interpretations of balance in the presence of relations with strengths, we need to have a measurement of relations. While it may appear arguable whether the relation strengths can be expressed by numerical values, it is fairly clear that the strength of any two relations can be compared. For positive relations such

as liking, valuing or approving, two relationships are comparable in terms of which one is stronger than the other. Similar argument applies to two negative relations. Finally, a positive relation and a negative relation are comparable by their signs. Hence, relations with strengths by nature inherit a total ordering.

We pick the ordering \preceq such that, $e_1 \preceq e_2$ denotes e_1 is equivalently or more positive than e_2 . In the simplest case where we have only positive and negative relations, we have that $+$ \preceq $-$. Following the definition of ordering \preceq , it is clear that for any $e_+ \in P$, $e_O \in O$, $e_- \in N$, $e_+ \preceq e_O \preceq e_-$ holds. We use $\langle e_1, e_2 \rangle$ to denote the set of relations $\langle e_1, e_2 \rangle = \{e \mid e_1 \preceq e \preceq e_2\}$. Hence, given $\langle e_1, e_2 \rangle$, the lower bound e_1 represents the strongest possible relationship and the upper bound represents e_2 represents the weakest possible relationship in this range.

Consider a single triad (A, B, C) , the tolerance of (A, B) is of the form $\langle e_1, e_2 \rangle$, a set of bounded values. That is, the range of comfortable relationships of (A, B) in (A, B, C) is constrained by an upper bound and a lower bound. As (A, B) 's tolerance is constrained by all the triads (A, B) is part of, it is the intersection of all individual tolerance and hence is also of the form $\langle e_1, e_2 \rangle$. Moreover, the two tolerance principles can be interpreted as the following.

Principle 3 (Transitivity of positive relationships.) *Let individuals A, B, C in a network form a triad, and (A, B) , $(A, B)'$, (B, C) be positive. Suppose $T = \langle e_1, e_2 \rangle$ denotes the tolerance of (A, C) based on relations $(A, B), (B, C)$, and $T' = \langle e'_1, e'_2 \rangle$ denotes the tolerance based on $(A, B)', (B, C)$. If $(A, B)' \preceq (A, B)$ then we have that $e'_2 \preceq e_2$.*

That is, if the relation between A, B is changed to a more positive relation $(A, B)'$, the corresponding tolerance of (A, C) will have a smaller upper bound. Equivalently, (A, C) 's tolerance is restricted to more positive relationships.

Principle 4 (Heterophily in relationships.) *Let individuals A, B, C in a network form a triad. Suppose $T = \langle e_1, e_2 \rangle$ denotes the tolerance of (A, C) based on relations $(A, B), (B, C)$, and $T' = \langle e'_1, e'_2 \rangle$ denotes the tolerance based on $(A, B)', (B, C)$. We have that if $(A, B)' \preceq (A, B) \preceq (B, C)$, or $(B, C) \preceq (A, B) \preceq (A, B)'$, then $e_1 \preceq e'_1$.*

That is, if the difference between (A, B) and (B, C) becomes larger, the corresponding tolerance of (A, C) will have a larger lower bound. Equivalently, (A, C) 's tolerance is restricted to more negative relationships.

Table 3.1: Tolerance rules in structural balance theory.

(A, B)	(A, C)	Tolerance for (B, C)	
+	+	$\langle +, O \rangle$	Transitivity
+	-	$\langle O, - \rangle$	Heterophily
-	-	$\langle +, - \rangle$	No stress

Clearly, the binary classification of relations in classic balance theory inherits a total ordering. We give the tolerance rules for Davis's WSBT in the general case of incomplete networks in Table 3.1 [4]. Notice that the table has neutral relationships " O ", which is not considered in their early work. This is because in the setting of incomplete networks, triads " $+, +, O$ " and " $+, -, O$ " are considered balanced implicitly in Davis's theory; if we substitute " O " with "+" in triad " $+, +, O$ " and substitute " O " with "-" in triad " $+, -, O$ ", we can get balanced triads in WSBT [1]. It is easy to verify tolerance rules in Table 3.1 agrees with Principle of Transitivity and Heterophily. According to the table, triads (1), (2) and (3) from Figure 1.1 are balanced as each relation is within the tolerance, but triad (4) is not balanced. In this aspect, our theory is a generalization of the Davis's WSBT.

Table 3.2: Five types of relations and their interpretations, displayed in the ordering, i.e. $s+ \preceq w+ \preceq O \preceq w- \preceq s-$.

Relation Type	Interpretation
Strongly positive (s+)	close friendship, trust
Weakly positive (w+)	aquaintance
Neutral (O)	unbiased relation, no relation
Weakly negative (w-)	minor disagreement, negative bias
Strongly negative (s-)	hatred, distrust

For a more detailed representation of relationships, consider a set of discrete classification of relationships shown in Table 3.2. This classification considers rela-

tion ties that have been discussed in previous literature. We show how to interpret balance in a more sophisticated manner in networks with these classes.

Strong positive ties, s+ are similar to close friendships. There is a strong expectation of reciprocity, similarity of tastes (homophily), common intentions and benevolence towards each other [19]. The traditional definition of SBT is based on these types of positive relationships.

Strong negative ties, s- are generally explained as having negative experiences with someone which is indicative of their negative intentions, unreliability and overall belonging to groups that are not considered trustworthy [20].

Table 3.3: Tolerance rules with strong, weak and neutral relations.

(A, B)	(A, C)	(B, C) 's ance	toler-	(A, B)	(A, C)	(B, C) 's ance	toler-
$s+$	$s+$	$\langle s+, w+ \rangle$		$w+$	$s-$	$\langle O, s- \rangle$	
$s+$	$w+$	$\langle s+, O \rangle$		O	O	$\langle s+, s- \rangle$	
$s+$	O	$\langle s+, w- \rangle$		O	$w-$	$\langle s+, s- \rangle$	
$s+$	$w-$	$\langle O, s- \rangle$		O	$s-$	$\langle w+, s- \rangle$	
$s+$	$s-$	$\langle w-, s- \rangle$		$w-$	$w-$	$\langle s+, s- \rangle$	
$w+$	$w+$	$\langle s+, w- \rangle$		$w-$	$s-$	$\langle s+, s- \rangle$	
$w+$	O	$\langle s+, s- \rangle$		$s-$	$s-$	$\langle s+, s- \rangle$	
$w+$	$w-$	$\langle w+, s- \rangle$					

However, these are not all the different classes of relationships that one might consider in a network. Granovetter [8] uses the term weak tie to refer to a relationship that is an acquaintance, not a close friend. Weak ties give access to less privileged information than strong ties, but come from outside of one's close network. In both cases, there is a tie between two people, but this does not imply a continuous interaction or a strong affective component as in trust.

We also introduce **weak negative ties, w-** to model cases in which there is a certain amount of distrust as a result of biases stemming from social groups people belong to, or heresay that may not be as strong as distrust [21]. In essence, the burden of proof of one's trustworthiness is much higher in weak distrust than in distrust, but in both cases, positive evidence is not evaluated in the same way as in trusting relations. These five types of relationships are summarized in Table 3.2.

The tolerance rules for networks with relations in Table 3.2 are given in Table 3.3. The rules are proposed according to many early work regarding strong and weak ties, as well as common observations in real life. The resulting imbalanced triads or structures are listed in Table 3.4. Notice that the triads with two positive relations and one negative relation are imbalanced as they are in classic balance theory, except for the ones in which all three relations are weak. In fact, the types of triads that consist of weak relations and neutral relations only are not considered to be imbalanced structures. Our argument here is that when all relations are weak or neutral, the influence inside the triad is not significant enough to draw tension. Also, triads of type “ $s+ s+ O$ ” and “ $s+ s- O$ ” are considered to be imbalanced structures. The arguments against each type of imbalanced structure is listed in Table 3.4.

Table 3.4: Imbalanced triadic structures in the presence of strong and weak ties, and arguments for the stress in the relation.

Triad	Argument for stress
$s + s + s-, s + s + w-,$ $s + w + s-, s + w + s-,$ $w + w + s-$	my two friends cannot get along with each other
$s + s + O$	my two close friends do not friend each other
$s + s - O$	my enemy’s close friend does not pick a side

3.3 Relation Distance and General Expression of Balance

The concept of extended balance is meaningful only if the tolerance rules can be explicitly defined, so that whether a triad and a network is balanced or not can be determined. Whenever relations are drawn from a finite ordered set, this is easy to do. However, as Cartwright and Harary also suggested, it is considerably more complex in the general case when the strengths of relations are drawn from arbitrary numerical values [3].

As a first step, we refine the measurement of relations with strengths from total ordering to positive real values. In particular, we define function $\psi : E \rightarrow R^+$

such that, for two relations $e_1, e_2 \in E$, $\psi(e_1) \leq \psi(e_2)$ if and only if $e_1 \preceq e_2$. Since positive values can be seen as metric distances, we call $\psi(e)$ the relation distance of e . In other words, relations with varying strengths are represented by distances with different lengths. More negative strengths are represented by larger distances and more positive strengths are represented by smaller distances.

We propose the following general rule of tolerance with the concept of relation distance.

Definition 2 *Let (A, B, C) be a triad of interest. The tolerance of (A, C) is given by $[|\psi(A, B) - \psi(B, C)|, \psi(A, B) + \psi(B, C)]$.*

It can be easily checked that the general tolerance rule agrees with Principle 3 and 4 by substituting \leq for \preceq . Immediately, we have the following theorem.

Theorem 2 *Given a triad (A, B, C) , if $\psi(A, B)$, $\psi(B, C)$, $\psi(A, C)$ satisfies the metric triangle inequality, then (A, B, C) is balanced.*

We first show that Table 3.1 and Table 3.3 can be considered as special cases of the general tolerance rule, under a mapping of relation strengths to distances. For Davis' WSBT and its tolerance rules in Table 3.1, we consider two thresholds: $b_+ < b_-$ such that if $\psi(e) \geq b_-$ then the relationship e is negative. Similarly, if $\psi(e) \leq b_+$, then e is a positive relationship. For any value $b_+ < \psi(e) < b_-$, the relationship is considered neutral. We can see that as long as $b_- > 2b_+$, Table 3.1 is equivalent to Definition 2.

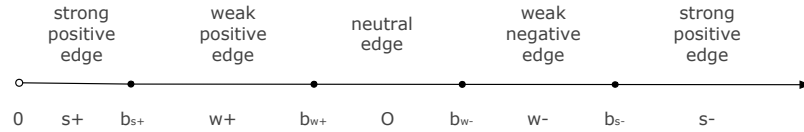


Figure 3.1: Partitioning of distance domain by boundary parameters $\{b_{s+}, b_{w+}, b_{w-}, b_{s-}\}$: distances within $(0, b_{s+}]$ are strong positive; (b_{s+}, b_{w+}) weak positive, $[b_{w+}, b_{w-}]$ neutral, (b_{w-}, b_{s-}) weak negative, and $[b_{s-}, \infty)$ strong negative.

Similarly, to capture the tolerance rules in Table 3.3, we consider the partitioning of the distance domain given in Figure 3.1. We can see that if the following conditions are satisfied for the boundary parameters:

Table 3.5: Constraints on the boundary parameters.

$$\begin{array}{ll}
b_{w+} > 2b_{s+} & b_{s-} > b_{w-} + b_{s+} \\
b_{s-} > 2b_{w+} & b_{w-} > b_{w+} + b_{s+}
\end{array}$$

then the tolerance rules given in Table 3.3 are equivalent to Definition 2, and all triads shown in Table 3.4 are imbalanced according to metric triangle inequality.

With the concept of relation distance, we are able to express the structure of a social network by drawing it in the metric space. The strength of each relation is expressed by the distance between two of its endpoints. Notice that every layout in the metric space automatically satisfies the metric triangle inequality, and hence corresponds to a balanced state of the network. We have the following the alternative definition of a balanced network.

Definition 3 *Given a social network G , G is balanced if it can be drawn in metric space such that each edge in G has the length of its relation distance.*

CHAPTER 4

CONVERGENCE MODEL

4.1 Convergence Model

Researchers have long argued that every social network has a tendency towards balanced states [7]. The balance aspect of “network evolution” becomes a non-trivial problem when relationships are measured numerically. Given an initially imbalanced network, it is interesting to see what state the network will converge to. We propose a model that characterizes a social network’s convergence in this chapter.

It is noted in social psychology literature that people are reluctant to make changes in relations as they tend to avoid the effort needed to make such changes. In a balanced triadic relation, participants are likely to do nothing and keep their pairwise relations as what they were. In an imbalanced triadic relation, participants are likely to make the smallest effort possible to regain triadic balance. We define the concept of relation cost as the effort one needs to take to accomplish a certain relation change. Our convergence model is established based on a unified assumption: every social network converges in a way that requires as little total change in relations as possible to reach a balanced state.

By definition 3, every layout in the Euclidean space automatically satisfies the metric triangle inequality, and hence corresponds to a balanced state of the network. For an imbalanced social network, it is not possible to draw it using its initial relation distances. Hence, our convergence model aims to produce a layout of the social network with minimum total relation cost from the original one.

Let $G = (V, E)$ denote an arbitrary social network, and $G^* = (V, E^*)$ denote a balanced state of G . Let $n * n$ matrix X denote the layout of G^* , with each row vector x_i denoting node i ’s location in m -dimensional space. For each pair (i, j) , $\psi(i, j)$ denotes its relation distance in G , and $d_{i,j}(X)$ denotes distance between i and j in X , i.e., its relation distance in G^* . Given an edge $(i, j) \in E$, we consider

the relation cost on (i, j) is given by:

$$c_{i,j}(X) = w_{\psi(i,j)} * (d_{i,j}(X) - \psi(i, j))^2$$

where the weight value is a function of the original distance. The weight function can take into account the difficulty of changing a relation. For example, it is generally easier to change a neutral relation than a positive or negative relation that carries with them initial bias. The study of optimal weights is beyond the scope of this thesis. However, we consider three main classes of weights:

$$w_{\psi(i,j)} = \begin{cases} w_+ & \text{if } \psi(i, j) \text{ is a positive edge} \\ w_O & \text{if } \psi(i, j) \text{ is a neutral edge} \\ w_- & \text{if } \psi(i, j) \text{ is a negative edge} \end{cases}$$

If $w_O \ll w_+$ and $w_O \ll w_-$, then neutral edges would have very little influence on the already established positive/negative relations.

Definition 4 Let $G = (V, E)$ be a social network where E is a set of weighted edges. Its converged network $G^* = (V, E^*)$ is given by layout matrix X with $d_{i,j}(X)$ as the relation distance between every pair (i, j) , such that the total relation cost $\sigma(X)$ is minimized,

$$\sigma(X) = \min_X \sum_{i < j \leq n} w_{\psi(i,j)} * (d_{i,j}(X) - \psi(i, j))^2$$

The optimization of relation cost is in fact a Metric Multidimensional Scaling problem (MDS) by assigning nodes a location in metric space. The total cost function is called stress in MDS, and is often minimized through an optimization strategy called *stress majorization* [10]. Stress majorization is an iterative method that guarantees monotonically decreasing stress in each iteration, and returns a locally minimum solution. It is recognized as a principled technique in the field of graph drawing. The algorithm, however, requires $O(n^3)$ time and $O(n^2)$ space. Due to its complexity, stress majorization is applicable on graphs with limited size. We discuss the algorithmic problem of stress majorization in Chapter 5.

4.2 Predictability of Convergence Model

A successful convergence model will help justify the balance tendency of social networks, and hence support the balance theory behind it. Moreover, convergence model inherits a predictive power as it models the dynamic nature of social networks. For example, the strong triadic closure states that two people with a common close friend are very likely to friend each other in the future [8]. We have argued in Table 3.4 that triad “ $s+, s+, O$ ” is imbalanced, it is important to verify if such a triad will converge to “ $s+, s+, +$ ” accordingly in the convergence model. Some frequently seen dynamic phenomenon in social networks are listed below, which we expect a congruent convergence model will be able to explain. The concept of strong and weak ties follows what is described in Table 3.2. Following the discussion in section 3.3, we partition the relation distance by the same constraints in Table 3.5. We illustrate and prove these propositions under our convergence model.

In the demonstration examples shown in Figure 4.1, Figure 4.2 and Figure 4.3, the boundary parameters are chosen as the following: $b_{s+} = 0.1$, $b_{w+} = 0.5$, $b_{w-} = 0.7$, $b_{s-} = 1.1$. The relation distances of strongly positive relation, weakly positive relation, neutral relation, weakly negative relation, strongly negative relation are picked as 0.1, 0.3, 0.6, 0.9, 1.1 respectively. Notice that the parameters is arbitrarily picked for the purpose of demonstration. Other configurations are legitimate as long as the constraints in Table 3.5 are satisfied. As it is discussed previously, we constrain w_0 such that $w_0 \ll \min \{w_+, w_-\}$. In the demonstration examples, the weights are chosen as: $w_+ = w_- = 1$, $w_0 = 0.001$.

In the proofs below, we denote G as the initial network and \mathcal{G} as its balanced state after convergence. The initial distance of a neutral relation is denoted as δ_0 . Clearly, $b_{w+} \leq \delta_0 \leq b_{w-}$ holds.

Proposition 1 *Strong Triadic Closure Property. Let (A, B, C) be a triad of interest. If both (A, B) and (B, C) are strongly positive relations in a social network, then (A, C) will be a positive relation in the balanced state.*

Illustrated in (1) of Figure 4.1, initially edges $(0, 1)$ and $(1, 2)$ are given as strongly positive. In its converged layout, node 0, node 1 and node 2 are on the same line

with node 1 in the middle. Hence, d_{02} is twice as much as d_{01} (d_{12}). As a result, $(0, 2)$ turns to be a weakly positive edge as desired.

Proof. We need to prove that the initially unconnected (A, C) will become a positive edge in \mathcal{G} . Also, we need to show (A, B) and (B, C) will remain strongly positive in \mathcal{G} .

Initially, $\psi(A, B) + \psi(B, C) < 2b_{s+} < b_{w+} < \delta_0 = \psi(A, C)$. A, B, C cannot coexist in the space without change some of the distances. To reach a balanced structure, it is either going to stretch (A, B) and (B, C) , or to shrink (A, C) in space. But in either way or a combination of both, the least modification, which indicates the smallest relation cost, occurs when A, B, C stays in a line with B in the middle. Let the length of both (A, B) and (B, C) be l_{1+} and l_{2+} , and the length of (A, C) be l_0 . The total relation cost should be the following.

$$\sigma(l_{1+}, l_{2+}, l_0) = \min_{l_{1+}, l_{2+}, l_0} w_0(l_0 - \delta_0)^2 + w_+(l_{1+} - \psi(A, B))^2 + w_+(l_{2+} - \psi(B, C))^2$$

subject to $l_0 = l_{1+} + l_{2+}$.

The optimization problem can be rewritten as the following,

$$\sigma(X) = \min_X 1/2 X^\top H X + C^\top X + \alpha,$$

where α is a constant and

$$X = \begin{pmatrix} l_{1+} \\ l_{2+} \end{pmatrix}, H = \begin{pmatrix} 2(w_0 + w_+) & 2w_0 \\ 2w_0 & 2(w_0 + w_+) \end{pmatrix}, C = \begin{pmatrix} -2(w_0\delta_0 + w_+\psi(A, B)) \\ -2(w_0\delta_0 + w_+\psi(B, C)) \end{pmatrix}.$$

Let $\frac{d}{dX}(\sigma(X)) = 0$. We have $X = -H^{-1}C$,

$$X = \begin{cases} l_{1+} = \frac{w_0\delta_0 + (w_0 + w_+)\psi(A, B) - w_0\psi(B, C)}{2w_0 + w_+} \\ l_{2+} = \frac{w_0\delta_0 + (w_0 + w_+)\psi(B, C) - w_0\psi(A, B)}{2w_0 + w_+} \end{cases},$$

and

$$l_0 = l_{1+} + l_{2+} = \frac{2w_0\delta_0 + w_+(\psi(A, B) + \psi(B, C))}{2w_0 + w_+}.$$

By constraint $w_0 \ll w_+$, we have

$$l_{1+} = \psi(A, B) + \frac{w_0(\delta_0 - \psi(A, B) - \psi(B, C))}{2w_0 + w_+} \simeq \psi(A, B)$$

Similarly, $l_{2+} \simeq \psi(B, C)$. Hence, both (A, B) and (B, C) will remain strongly positive. Now, $l_0 \simeq \psi(A, B) + \psi(B, C) < b_{w+}$, and therefore (A, C) will become a positive relationship. \square

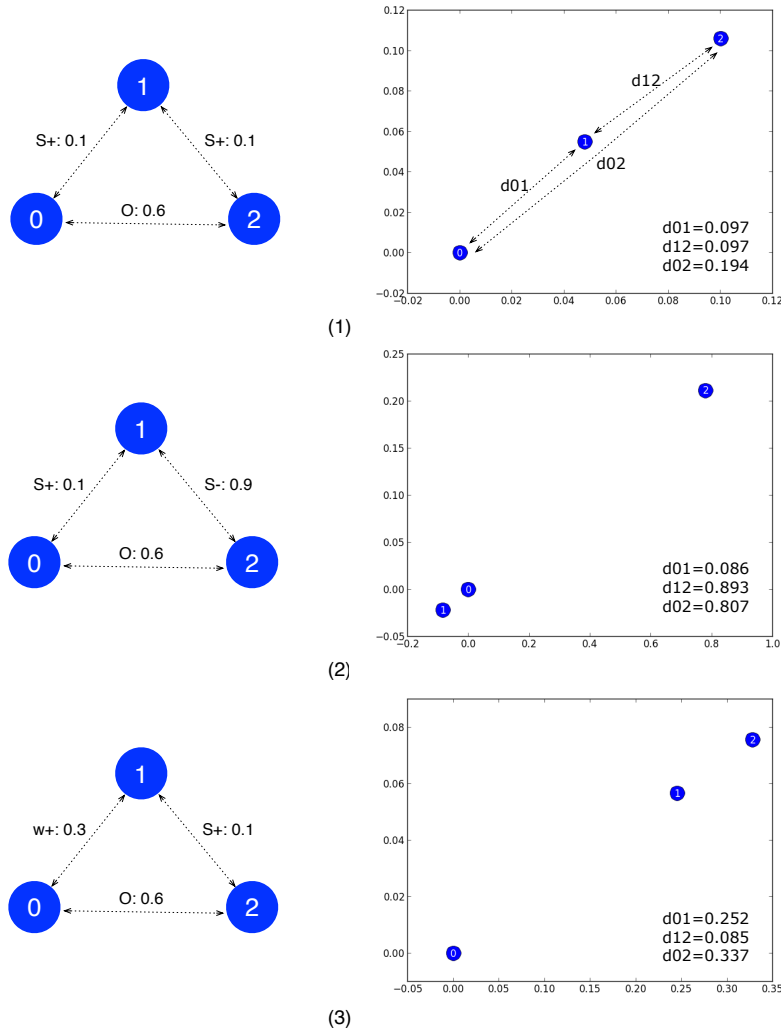


Figure 4.1: The demonstration experiments of Propositions 1, 2 and 3. Each blue ball has a white number i in it, indicating node i . The initial state of each triad is shown in figures on the left, and the corresponding converged layouts on the right.

Proposition 2 *Neutralization Property.* Let (A, B, C) be a triad of interest. If (A, B) is strongly positive and (B, C) is strongly negative, then (A, C) will be negative in the balanced network, and the strength of (A, C) should be no stronger than the strength of (B, C) .

The first part of the neutralization property can be seen as a twin property of strong triadic closure with negative sign. It says if A has a friend B and a strong enemy C, then B and C will not get along in the long run. The second part is an inference assumption; one argument for the assumption is: “the later established disliking is not likely to be stronger than the original hatred that causes the disliking”.

Illustrated in (2) of Figure 4.1, initially edge $(0, 1)$ is given as strongly positive while edge $(1, 2)$ is given as strongly negative. In its balanced layout, node 0, node 1 and node 2 are on the same line with node 0 in the middle. Hence, d_{02} equals $d_{12} - d_{01}$, which is in the range of negative relations.

Proof. We need to prove (A, C) will become a negative edge in \mathcal{G} , and that (A, C) will not be longer than (B, C) in metric space. Also, we need to prove (A, B) and (B, C) will remain what they were in \mathcal{G} .

By similar arguments in the proof of Proposition 1, A, B, C cannot coexist in the space without change some of the distances. To reach a balanced structure, it is either going to stretch (A, B) and (A, C) , or to shrink (B, C) in space. But in either way or a combination of both, the least modification, which indicates the smallest total relation cost, occurs when A, B, C stays in a line with A in the middle. Let the length of (A, B) be l_+ , the length of (B, C) be l_- and the length of (A, C) be l_0 . The total relation cost on (A, B, C) during the convergence should be the following.

$$\sigma(l_+, l_0, l_-) = \min_{l_+, l_0, l_-} w_0(l_0 - \delta_0)^2 + w_+(l_+ - \psi(A, B))^2 + w_-(l_- - \psi(B, C))^2$$

$$\text{subject to } l_0 + l_+ = l_-.$$

The optimization problem can be rewritten as the following,

$$\sigma(X) = \min_X 1/2X^\top HX + C^\top X + \alpha,$$

where α is a constant and

$$X = \begin{pmatrix} l_0 \\ l_+ \end{pmatrix}, H = \begin{pmatrix} 2(w_0 + w_-) & 2w_- \\ 2w_- & w(w_+ + w_-) \end{pmatrix}, C = \begin{pmatrix} -2(w_0\delta_0 + \psi(A, C)) \\ -2(w_+\psi(A, B) + \psi(B, C)) \end{pmatrix}.$$

Let $\frac{d}{dX}(\sigma(X)) = 0$. We have $X = -H^{-1}C$,

$$X = \begin{cases} l_0 = \frac{(w_+w_0+w_0w_-)\delta_0+w_+\psi(A,C)-w_+w_-\psi(A,B)}{w_0w_++w_+w_-+w_0w_-} \\ l_+ = \frac{(w_+w_0+w_+w_-)\psi(A,B)+w_0\psi(A,C)-w_0w_-\delta_0}{w_0w_++w_+w_-+w_0w_-} \end{cases},$$

and

$$l_- = \frac{w_+w_0\psi(A, B) + w_+w_0\delta_0 + (w_+ + w_0)\psi(A, C)}{w_0w_+ + w_+w_- + w_0w_-}.$$

By constraints $w_0 \ll \min\{w_+, w_-\}$ and $w_- = 1$, we have

$$l_0 \simeq \frac{w_+\psi(A, C) - w_+w_-\psi(A, B)}{w_+w_-} = \frac{\psi(A, C)}{w_-} - \psi(A, B) = \psi(A, C) - \psi(A, B),$$

$$l_+ \simeq \psi(A, B),$$

$$l_- \simeq \frac{\psi(A, C)}{w_-} = \psi(A, C).$$

Hence, (A, B) will remain strongly positive, (B, C) will remain strongly negative and (A, C) will be no longer than (B, C) . Since $l_0 \simeq \psi(A, C) - \psi(A, B) > b_{s-} - b_{s+} > b_{w-}$, (B, C) will be a negative relation after convergence. \square

Proposition 3 *Attenuation Property. Let (A, B, C) be a triad of interest. If (A, B) and (B, C) are both positive, then (A, C) will be non-negative in balanced network, and the strength of (A, C) should be no stronger than the strength of either (A, B) or (B, C) .*

The first part of the propagation attenuation property can be seen as a weaker triadic closure property. It says that if two people have a common friend, then based on this information alone, they cannot be enemies. The second part, a similar inference assumption as the one in Proposition 2, can be explained by “the later

established liking is not likely to be stronger than the original brotherhood that invokes the liking”.

Illustrated in (3) of Figure 4.1, initially edge $(0, 1)$ is weakly positive while edge $(1, 2)$ is strongly positive. Similar to the situation in (1), $d_{02} = d_{01} + d_{12}$, which is in the range of weakly positive relations. Also, of course, $d_{02} > \max\{d_{01}, d_{12}\}$, which agrees with the attenuation property.

Proof. Let (A, B) , (B, C) be positive edges. Suppose that $\psi(A, B)$, $\psi(B, C)$, $\psi(A, C)$ still violate the triangle inequality. By the same arguments in the proof of Proposition 1, we still conclude that in the balanced network (A, C) will be a positive edge, with its strength no stronger than the strength of either (A, B) or (B, C) . Otherwise, since (A, B) , (B, C) , (A, C) are able to coexist with its original relation distance, all of the three edges will remain what they were in the balanced network with a total relation cost of 0. Hence, (A, C) will remain as a neutral edge. \square

By Proposition 2 and Proposition 3, we further have the following two corollaries regarding positive and negative paths.

Corollary 1 *Positive Path.* *Given a path that consists of positive edges only, the positive relation will propagate with its strength being attenuated along the path, until it becomes a neutral relation.*

Let A_1, A_2, \dots, A_n be a positive path, such that every (A_1, A_2) is a positive relation. The corollary states that every (A_1, A_k) will be a non-negative relation, and that (A_1, A_{k-1}) will be positively stronger than (A_1, A_k) .

Corollary 2 *Negative Path.* *Given a path that consists of only one negative edge, the relation between the end nodes of the negative path is non-positive in the balanced network.*

We demonstrate the transitivity and attenuation along a Positive Path in (1) (2) (3) of Figure 4.2. (1) of Figure 4.2 shows the converged layout of a positive path consisting of 6 strongly positive relations. In this configuration, since the sum of all 6 distances is smaller than the relation distance of a neutral edge, all nodes stay on the same line. Clearly, $d_{06} > d_{05} > \dots > d_{02}$, which agrees with the attenuation

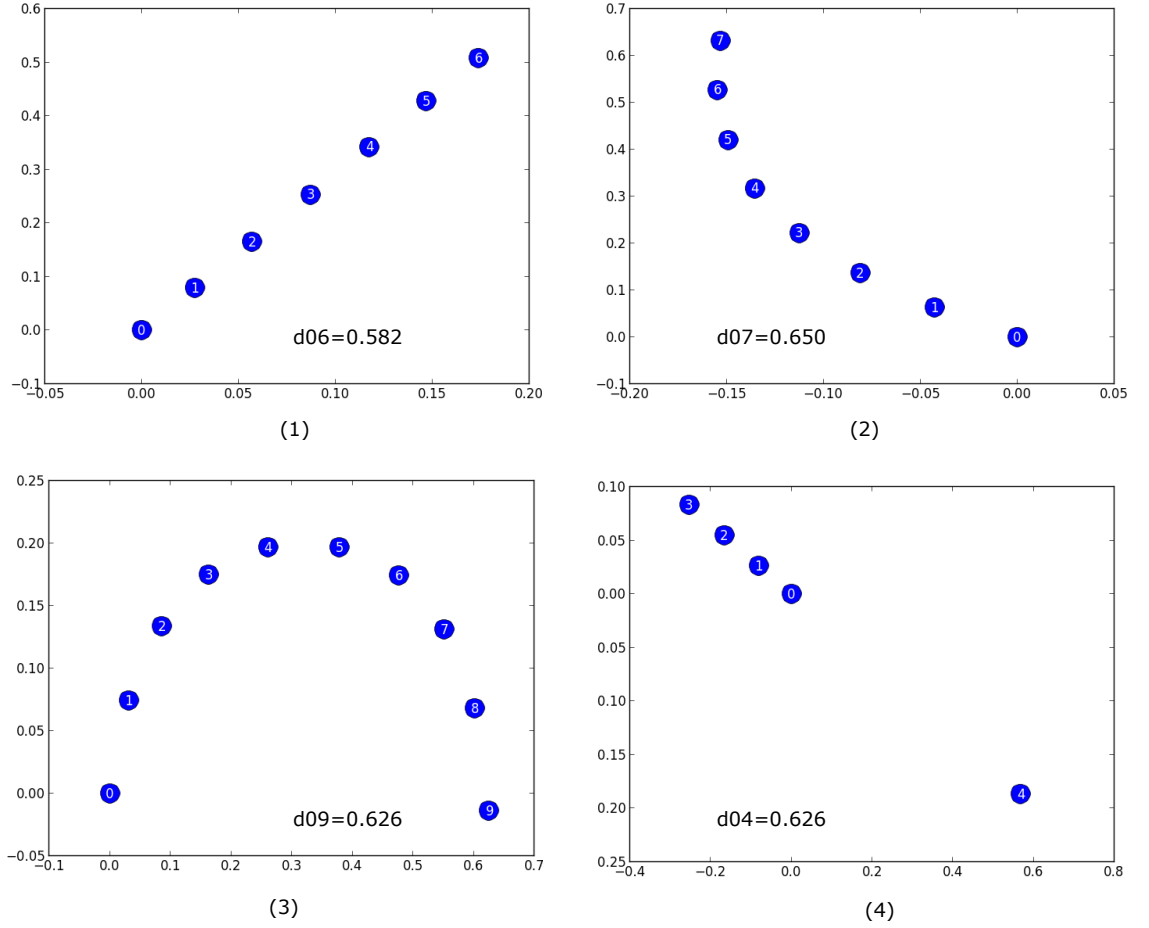


Figure 4.2: The demonstration experiments of Corollary 1 and 2. Each blue ball has a white number i in it, indicating node i 's location in its converged layout.

property. d_{06} turns to be in the range of neutral relations, which indicating the relationship between node 0 and node 6 becomes neutral.

(2) (3) of Figure 4.2 show the converged layout of a positive path consisting of 7 and 9 strongly positive relations respectively. In such cases, the sum of all 7 or 9 distances becomes larger than the relation distance of a neutral edge. As a result, the layout of nodes curves. Both d_{07} and d_{09} are in the range of neutral relations. Moreover, d_{07} is consistently larger than d_{09} in our experiments, indicating the distance between two endpoints of the positive path is non-increasing w.r.t. the path length. The empirical results of (2)(3) show the non-negativeness of the relationship between the endpoints of a positive path.

(4) of Figure 4.2 is the converged layout of a negative path, in which (3, 4) is the only negative edge. Similar to positive path, it has a non-positive relationship between its two endpoints.

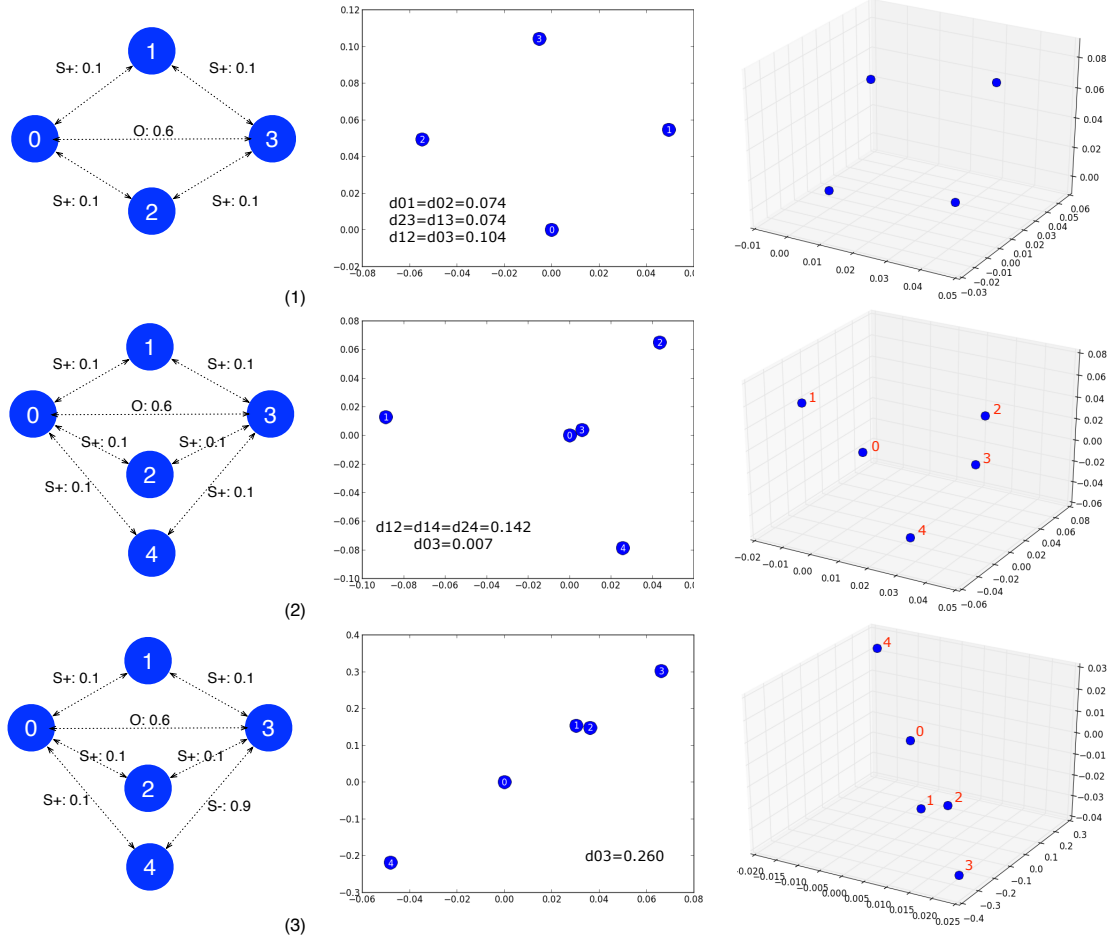


Figure 4.3: The demonstration experiments of Proposition of Aggregation. Each blue ball has a white number i in it, indicating node i . The initial state of each triad is shown in figures on the left, the corresponding converged layouts in 2-D space is shown in the middle, and the corresponding converged layouts in 3-D space is shown on the right.

Proposition 4 *Aggregation Property.* Suppose there are multiple positive and negative paths between two nodes, then adding one more positive path between them will make their relation more positive while adding one more negative path will make the relation more negative.

Aggregation property is a generalization of “people with more common friends are more likely to become friends”. It says if two people A,B have more (direct or indirect) friends in common, then they are more likely to be friends, or less likely to be enemies. On the other hand, if they have more conflicting people in common, who are (direct or indirect) friends of A and (direct or indirect) enemies of B, then they are more likely to be enemies, or less likely to be friends.

To illustrate the Property of Aggregation, we run 3 different tests. In the first experiment shown in (1) of Figure 4.3, edges $(0, 1)$, $(0, 2)$, $(1, 3)$, $(2, 3)$ are given as strongly positive initially. The converged layout forms a square, and $d_{03} = \sqrt{2}d_{01}$. Obviously, d_{03} is in the range of positive relations. Moreover, d_{03} is smaller than d_{02} in (1) of Figure 4.1, which indicates adding a common friend between two people will grant a better chance to them to become friends. In the second experiment shown in (2) of Figure 4.3, we add another positive path $0, 4, 2$. The converged layout is still symmetric, but the distance between node 0 and node 3 becomes very close, which agrees with the concept of aggregation. In the third experiment shown in (3) of Figure 4.3, we add a negative path instead. In the converged layout, the distance between node 0 and node 3 becomes larger than d_{02} in (1) of Figure 4.1, which indicating adding a conflicted common neighbor between two people will deteriorate their relationship.

When the number of nodes is more than 3, a two-dimensional drawing is not necessarily able to express the layout of minimum relation cost. For example, in (2) of Figure 4.3, d_{03} would be larger if drawing in a 3-D space. Due to the limit of dimensionality, node 0 and node 3 are forced to squeeze together. In studying the effects of dimensionality, we repeat the same aggregation demonstrations in 3-D space. In (1) of Figure 4.3, the layout still forms a square. In (2) of Figure 4.3, the layout forms a triangular bipyramid, and the distance between node 0 and node 3 becomes reasonably small, but is still smaller than it is in (1) of Figure 4.3. In (3) of Figure 4.3, the distance between node 0 and node 3 is slightly larger than it is in (1) of Figure 4.3, showing a more reasonably deterioration of relationship by adding a negative path.

Proof. We present the general idea of the proof. See appendix for the full

proof. Initially, strongly positive relations (A, B) , (B, D) are given. By Proposition 1, the length of (A, D) will be approximately $\psi(A, B) + \psi(B, D)$ in the balanced network.

Suppose one more positive paths is added between A and D . For example, strong positive edges (A, C) and (C, D) are added. Neutral edge (B, C) needs to stretch out to reach their initial length b_{w+} . But if edge (B, C) stretches more, edge (A, D) has to stretch less in space. In the balanced network \mathcal{G} , A, B, C, D will form a quadrilateral in the Euclidean space. Hence, the length of (A, D) will be smaller than $\psi(A, B) + \psi(B, D)$.

Moreover, with more neutral edges like (B, C) on the multi-plane that is vertical to (A, D) , the network structure will be balanced at a point where (A, D) is less stretched, so that the total relation cost is minimized.

On the other hand, if adding one more negative path between A and D , then equivalently a non-positive relation will be added between them according to corollary 2. If it is a negative relation, (A, D) will be directly stretched out, because negative relation is longer than neutral relation in metric space. \square

CHAPTER 5

ALGORITHMS FOR STRESS MAJORIZATION

In this chapter, we review the problem of stress majorization in the MDS literature [10]. Stress majorization is recognized as a principled technique in graph drawing [22]. Due to its complexity, however, classic stress majorization is applicable on graphs with limited size. Modern social networks usually consist of over 10^5 nodes, which are typically infeasible for stress majorization. We discuss two algorithms that together give solution to our convergence model in the context of large social networks.

5.1 Review of stress majorization

Following Gansner's description, we denote the d -dimensional location matrix as X [10]. Let node i 's location be X_i . The MDS problem is to minimize the following stress function

$$stress(X) = \sum_{i < j} w_{ij} (||X_i - X_j|| - d_{ij})^2, \quad (5.1)$$

where w_{ij} are weights and d_{ij} denote the ideal distance between i and j . There are many ways to minimize $stress(X)$. Among them, De Leeuw's stress majorization technique is considered to be a major achievement on the subject. The iterative majorization method at each step minimizes a simple convex function, which guarantees a monotonically convergent solution [22]. In application, the majorization process involves solving the following equation iteratively. For a detailed analysis of Stress Majorization, please refer to Appendix A.

$$L^w X(t+1) = L^{X(t)} X(t), \quad (5.2)$$

where L^w and $L^{X(t)}$ are given by

$$L_{i,j}^w = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j \end{cases}, \quad L_{i,j}^{X(t)} = \begin{cases} -w_{ij}d_{ij}/\|X(t)_i - X(t)_j\| & i \neq j \\ -\sum_{j \neq i} L_{i,j}^{X(t)} & i = j \end{cases}.$$

At each iteration, L^w is constant throughout the process, whereas matrix $L^{X(t)}$ needs to be computed recursively. While stress majorization is an effective and accurate optimization strategy, it is not feasible in very large networks. Solving Equation 5.2 in each iteration requires the computation of a fully dense matrix, which in total requires $O(n^3)$ time and $O(n^2)$ space. For network of node number $> 10^5$, direct implementation of stress majorization becomes impossible.

5.2 Stress Majorization in Large Social Networks

We discuss two algorithms to solve the optimization given in Equation 5.1 in large social networks.

Algorithm 1. First, we note that social networks are usually sparse. In other words, there are a lot more “no-links” than “real links”. If we set these neutral edge’s weight w_O to 0, L^w will be a sparse matrix. Moreover, the right side of the iteration, $L^{X(t)}X(t)$, can be expanded to the following:

$$(L^{X(t)}X(t))_i = \sum_{j \neq i} w_{ij}d_{ij} \frac{X_i - X_j}{\|X_i - X_j\|}. \quad (5.3)$$

Hence, we will only need to sum up the terms corresponding to “real links” to get $L^{X(t)}X(t)$. Let the average degree of a social network be k , computing $L^{X(t)}X(t)$ only requires $O(kn)$. The cost to solve the linear equation 5.2 with sparse matrix L^w is usually much less than $(2/3n^3)$, depending on the patterns of L^w . As a result, the problem is likely to become feasible in large social networks.

However, we cannot simply we set w_O to 0, because otherwise relationship changes regarding initially neutral relations would add zero relation cost. As a result, arbitrary changes over initially neutral relations would occur, and properties such as transitivity of positive relationships would not necessarily hold anymore.

Our second observation is inspired by the famous small world phenomena [23],[24].

If two people are somehow connected by a path of relations, the path is often of very short length. In the setting of online social networks, we regard two people who are connected by paths of more than length 4 distant from each other, and hence assume their relationship will remain neutral over the convergence. As a result, we only need to consider neighbors of each node within two links, which is in average k^2 in number. Under this constraint, computing $L^{X(t)}X(t)$ requires $O(k^2n)$ while L^w is still sparse (k^2n non-zero entries). The convergence problem is solvable in large social networks while the principle of relation cost minimization is not violated.

Algorithm 2. The second algorithm is inspired by Khoury’s latest work. They

```

Data:  $G, w_+, w_-, w_O, \theta$ 
Result:  $X$ 
Create a sparse matrix  $L^w$ ;
for each node  $i$  in  $G$  do
    for each direct neighbor  $j$  of node  $i$  do
        if  $(i, j)$  is a positive edge then
            | set  $L_{ij}^w = w_+$ ;
        else
            | set  $L_{ij}^w = w_-$ ;
        end
    end
    for each 2-step neighbor  $k$  of node  $i$  do
        | set  $L_{ik}^w = w_O$ ;
    end
end
Initialize  $X'$  as a random layout;
repeat
     $X = X'$ ;
    compute  $stress(X)$  by Equation 5.1;
    compute  $X' = (L^w)^{-1}L^X X$  following Equation 5.2 and Equation 5.3;
    compute  $stress(X')$  by Equation 5.1;
until  $(stress(X) - stress(X'))/stress(X) < \theta$ ;
return  $X$ ;

```

Algorithm 1: Stress Majorization with sparse L^w

first proposed the problem of drawing very large graphs by low-rank stress majorization [25]. Our algorithm is a refined version of their work, designed for applications in large social networks. In the setting of a social network, X denotes the location

matrix of the each node, and w_{ij} and d_{ij} are attributes regarding social links. Many social networks, such as the ones with signed edges (positive/negative), have limited type of edges. Hence, we assume that both w_{ij} and d_{ij} have finite values. Precisely, each $w_{ij}d_{ij}$ has finite values. For example, in a social network with signed edges, w_{ij} and d_{ij} should have only two values corresponding to a positive edge and a negative edge. In graph drawing, $w_{ij}d_{ij} = 1$ is typically assumed [10], [25]. We do not have such an assumption here.

The high complexity of stress majorization comes from computing Equation 5.2 iteratively. The computation can be divided into two parts. 1. Solve equation $L^w X(t+1) = b$, assuming $b = L^{X(t)} X(t)$ is given; 2. Compute $L^{X(t)} X(t)$ efficiently.

For the first sub-problem, Khoury first proposed a low-rank SVD based approximation approach. Inspired by the work by Drineas, they used a random sampling technique to approximate L^w with only a small fraction of its columns [26]. Every matrix M admits a SVD decomposition of the form $M = U\Sigma V^T$, where U and T are orthogonal matrixes. When the diagonal elements of Σ are in non-increasing order, setting all but the first k values of Σ to zero yields a sequence of increasingly better approximation of M as k increases. If most values of Σ are relatively close to zero, zeroing them does not change M much. For matrix L^w , it satisfies equation $L^w = D^w + O^w$, where O^w is an off-diagonal matrix with each element $O_{ij}^w = -w_{ij}$ and $D^w = -\text{diag}(O^w \mathbf{1})$ is a diagonal matrix. Since matrix D^w depends on O^w and O^w has better k -rank approximation than original L^w , we approximate O^w instead. Consequently, L^w is represented by the following form

$$L^w = D^w + O^w = -\text{diag}(U\Sigma V^T \mathbf{1}) + U\Sigma V^T. \quad (5.4)$$

Our next step is to approximate O^w by the k most significant values of Σ . In other words, we will only need to store two $n \times k$ matrixes and a $k \times k$ matrix.

Since L^w is singular, $L^w X(t+1) = b$ is undetermined. We consider its minimum norm solution, $L^{w+} b$, where L^{w+} is the Moore-Penrose Pseudo-inverse of L^w [27]. Since matrix L^w is singular and that its null space is based by exactly $\mathbf{1}$, we project

L^w out to circumvent the singularity.

$$L^{w+} = (L^w + \mathbf{1}\mathbf{1}^T)^{-1} - \mathbf{1}\mathbf{1}^T \quad (5.5)$$

$$(L^w + \mathbf{1}\mathbf{1}^T)^{-1} = (U\Sigma V^T + D^w + \mathbf{1}\mathbf{1}^T)^{-1} \quad (5.6)$$

Let $A = D^w + \mathbf{1}\mathbf{1}^T$, and by the Woodbury Matrix Identity [28],

$$(L^w + \mathbf{1}\mathbf{1}^T)^{-1} = (A + U\Sigma V^T)^{-1} = A^{-1} - A^{-1}UT_1^{-1}V^TA^{-1}, \quad (5.7)$$

where

$$A^{-1} = (D^w)^{-1} - T_2T_3^{-1}T_2^T,$$

$$T_1 = (\Sigma^{-1} + V^TA^{-1}U),$$

$$T_2 = (D^w)^{-1}\mathbf{1},$$

$$T_3 = 1 + \langle \mathbf{1}, T_2 \rangle.$$

Since D^w can be written as $\text{diag}(U\Sigma V^T\mathbf{1})$, the above computation only involves matrix Σ , U and V . If only the k most significant values of Σ is taken, $L^{w+}b$ can be efficiently computed. Specifically, the computational cost is $O(k^3)$. The next question is how to accurately approximate the SVD of O^w .

In [26], the authors propose a randomized algorithm to approximate the SVD of a matrix by only sampling k columns of the matrix. In particular, each column M_i of a matrix M is sampled with a probability of $p_i = \|A_i\|^2 / \|A\|_F^2$. Each picked column is then scaled by a factor of $1/\sqrt{kp_i}$, the obtained $n \times k$ matrix \hat{M} is shown to approximate M with good error bound. The computation of SVD of such \hat{M} requires $O(nk^2)$ time.

Given the assumption that w_{ij} has limited number of values, we are able to compute the 2-norms without accessing every item of L^w . For example, in the social network with signed edges, for each node (column) i , its norm can be computed by counting the number of positive edges and negative edges it connects to, which takes $O(dn)$ where d is the average degree. In many applications, we expect the average degree is much smaller than n .

In total, we can see the total running time of the first part algorithm is $O(k^3 + nk^2 + nd)$, where k is a selected rank and d is the average degree of the network.

For the second sub-problem, the computation of $L^{X(t)}X(t)$ can be well approximated using Barnes-Hut algorithm of n-body problem with high efficiency [29]. As it is in Equation 5.3, $L^{X(t)}X(t)$ is a weighted sum of unit vector to i from every other nodes. This can be treated as an n-body problem, where every weighted vector is the force every node j puts on i . The brute force algorithm for such n-body problem requires $O(n^2)$ computations in total. Barnes and Hut developed an algorithm to accurately solve the n-body problem in $O(n \log n)$ [29].

The Barnes-Hut algorithm groups together bodies that are sufficiently nearby. It recursively divides the set of bodies into groups by storing them in a quad-tree. To calculate the net force on a particular body, we traverse the nodes of the tree, starting from the root. If the center-of-mass of an internal node is sufficiently far from the body, we approximate the bodies contained in that part of the tree as a single body, whose position is the groups center of mass and whose mass is the groups total mass. The algorithm is fast because we do not need to individually examine any of the bodies in the group. In graph drawing, the individual force of each member inside a distant group is the same. If the internal node is not sufficiently far from the body, we recursively traverse each of its subtrees. To determine if a node is sufficiently far away, we compute the quotient s/d , where s is the width of the region represented by the internal node, and d is the distance between the body and the nodes center-of-mass. Then, we compare this ratio against a threshold value θ . If $s/d < \theta$, then the internal node is sufficiently far away. By adjusting the θ parameter, we can change the speed and accuracy of the simulation.

Let the target body be i , and J is a group far from i . We denote a super-node at the center of J be j . The total force of J on i can be estimated by $|J|$ times the force from j to i . In our setting when $w_{ij}d_{ij}$ has limited number of values, the individual force of each member in J is not necessarily the same. However, J can be partitioned into finite subgroups J_1, J_2, \dots, J_P . For each subgroup, the same technique can be used to compute the overall force from each J_p to i . The Barnes-Hut algorithm reduces the complexity of computing the net force on a particular

body to $O(\log n)$. In our setting, it takes $O(P \log n)$ instead, where P is the number of values of $w_{ij}d_{ij}$. In total, computing $L^{X(t)}X(t)$ takes $O(Pn \log n)$ in time.

With the solutions to the two sub-problems, we are able to solve Equation 5.2 even on very large networks. The computational cost per iteration is in total $O(k^3 + n(k^2 + d) + Pn \log n)$, where k is the selected rank, d is the average degree of the network and P is the number of types of edges of the network.

Function `solveA-1t(t, Dw)`
 compute $T_2 = (D^w)^{-1}\mathbf{1}$;
 compute $T_3 = 1 + \langle \mathbf{1}, T_2 \rangle$;
return $A^{-1}(t) = (D^w)^{-1}t - T_2 T_3^{-1} T_2^T t$;

Function `solveLw+b(b, Dw, U, VT, Σ)`
 compute $v_1 = \text{solve}_{A^{-1}}t(b, D^w)$;
 compute $A^{-1}U = \text{solve}_{A^{-1}}t(U, D^w)$;
 compute $T_1 = \Sigma^{-1} + V^T A^{-1}U$;
 compute $v_2 = U(T_1)^{-1}V^T v_1$;
 compute $v_3 = \text{solve}_{A^{-1}}t(v_2, D^w)$;
return $L^{w+b} = v_1 - v_3$;

Data: $G, w_+, w_-, w_O, \theta, k$

Result: X

Approximate the SVD of L^w by sampling k columns of it following the method in [26];

Compute and store the corresponding U, V^T, Σ and D^w ;

Initialize X' as a random layout;

repeat

$X = X'$;
 compute $\text{stress}(X)$ by Equation 5.1;
 compute $L^X X$ through the Barnes-Hut algorithm;
 approximate $X' = \text{solve}_{L^{w+b}}(L^X X, D^w, U, V^T, \Sigma)$;
 compute $\text{stress}(X')$ by Equation 5.1;

until $(\text{stress}(X) - \text{stress}(X')) / \text{stress}(X) < \theta$;

return X ;

Algorithm 2: Stress Majorization based on Approximated SVD

CHAPTER 6

EXPERIMENTS AND RESULTS

Suppose we are given a social network with signs, but a small fraction of the edge signs are “hidden”. How can we predict these signs with the information provided by the rest of network? In this chapter, we study the *edge sign prediction problem* using online social network data. The prediction algorithm in this work is a direct implementation of the convergence model. Our experimental results match and surpass the state of art. The results provide empirical support of the convergence model, as well as the extended balance theory behind it.

We justify that the convergence model is able to predict these “hidden” signs from the theoretical aspect. Let’s denote the original social network with all signed edges as G , the network consisting of hidden edges as G_h , and the network consisting of the remaining edges as G_r . The edges (relations) between each pair of nodes is measured by $\{+, -, O\}$. We run the convergence model on G_r , and denote the network after convergence as G'_r . We expect that the signs of the hidden edges in G'_r largely agree with the true signs.

By the assumption that every social network has a tendency towards balance, it can be inferred that G is largely balanced at any moment. Hence, the majority of G_r is balanced. The only exceptions are the components with hidden edges, which are of sign O in G_r . By the principle of total relation cost minimization, the changes mostly occur on the O -sign hidden edges during the convergence. We expect the hidden edges in G_h to have their true signs in G'_r if G is largely balanced.

We compare our algorithm to force directed algorithm (FD) in [6]. Note that we have tuned our implementation of FD to provide similar performance reported in this work. Even though this work combines two algorithms, in our comparison experiments, we find that FD alone gives equally good prediction performance on all three datasets as the combination. Despite its more global nature, the second algorithm (PP) from [6] does not contribute to significant performance gains.

6.1 Data and Algorithm

We use the same three datasets as it is in [5], [6] to conduct our experiments, all provided by the Stanford Large Network Dataset Collection.

- Epinions is a product review website where users give reviews and ratings on product articles. Users can choose to trust or distrust others. The network contains more than 100,000 users and over 700,000 trust/distrust edges.
- Slashdot is a technology news website where users rate each other as friends or foes. The dataset released in February 2009 contains over 77,000 users and over 900,000 friend/foe edges.
- Wikipedia elections collects the votes by Wikipedia users in elections for promoting candidates as administrators. Each user can give a supporting (positive) or opposing (negative) vote on the promotion of another. The dataset has about 7,000 users and around 100,000 votes(edges).

All edges are treated as undirected in our experiments. Due to both the memory and computational cost of SM, running SM on the entire dataset is infeasible. As a result, we generate random samples of our datasets using the snowball sampling method in which a small number of seeds with degree greater than a given threshold deg are selected at random, then all nodes that are adjacent to the seed node are selected iteratively until the desired network size is reached. In our practice, the size of the resulting graph is in the range 3,000 – 5,000 nodes, k is chosen from 2 – 10 randomly and deg is chosen from 7 – 20 randomly. For each dataset, we generate 10 such sub-networks and run tests through 10-fold cross validation. The number of edges in a sub-network of Epinions is around 180,000, the number of *Slashdot* is around 65,000 and the number of Wikipedia votes is around 160,000. In the implementation of SM, the partitioning of the distance domain satisfies $b_+ < 1/2b_-$, conforming to our theory. The weight of each type of edge satisfies $w_O \ll w_+ < 1/2w_-$. The first inequality has been argued in previous section. The second one is chosen empirically, indicating that a negative edge has larger influence than a positive one. Note that we use the same setting for all the networks and do not employ any other adjustable parameters.

Data: M, k, deg, G
Result: Pt, Nt
 Get $G' = \text{generate-subgraph}(M, k, deg, G)$;
 Partition G' into 10 groups of test and training samples;
 Create two empty sets Pt, Nt ;
for *each of the groups* **do**
 run SM on the training sample and get the layout;
 for *each edge e in the testing sample* **do**
 compute its distance in the layout;
 if *e is a positive edge* **then**
 add its distance to Pt ;
 else
 add its distance to Nt ;
 end
 end
end

Algorithm 3: SM Prediction

6.2 Results

The distances of testing edges are computed by the layout of the training data. Given a distance threshold, the sign of each edge is predicted as positive if and only if its distance is smaller than the threshold. In the previous work, such threshold is computed from the (distance, sign) pairs of the training samples using standard machine learning techniques [5], [6]. In this thesis, however, we do not concentrate on the learning process. The issue of interest is how good the convergence model performs in separating hidden positive edges from negative ones in terms of distance. Instead of making predictions based on particular threshold, we draw ROC curves for evaluation. We consider different distance thresholds and compute the false and trust positive rates based on the computed Pt (Nt) values returned by the SM Prediction Algorithm. The ROC curves in Figure 6.1 are drawn upon the Pt (Nt) values from the accumulation of all testing samples.

For all three datasets, we find the ROC curve of SM is on the “northwest” side of the one of FD , which indicates SM is consistently better than FD in separating hidden positive edges from negative ones. Notice that the improvement for Slashdot is the most significant one among the three. The possible explanation is in Slashdot edges represent “friends” or “foes”, which is by nature a more clear identification

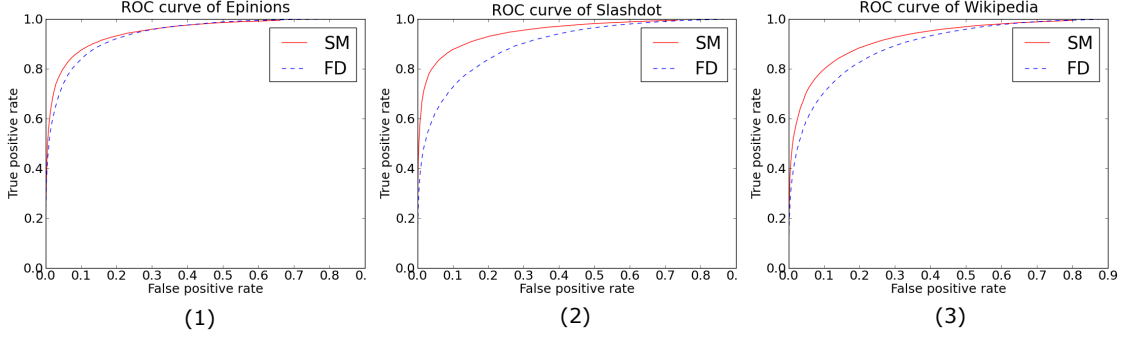


Figure 6.1: The ROC curves are drawn upon distances of hidden edges, generated by SM and FD. (1) (2) (3) correspond to the ROC curves generated from Epinions, Slashdot and Wikipedia datasets respectively.

of trust and distrust, whereas in votes in Wikipedia or distrusting reviews of other users in Epinions are not as strong distrust relations. As a result, our convergence model characterizes its mutual relation based structure nicely, and hence produces good prediction performance.

On the Epinions and Slashdot datasets, the best thresholds on ROC curve give 88 – 90% accuracy on both positive and negative hidden edges. For Wikipedia, we achieves 83 – 85% at the best threshold. The accuracy rates of Epinions and Wikipedia match the best results from previous work, and the one of slashdot appears to be the best so far.

6.3 General Edge Prediction

The *edge sign prediction* only deals with the existing signed edges. In predicting the sign of a hidden edge, we already know that the edge exists in the original network. A more general problem is to predict whether there is positive edge between arbitrary pair of nodes [18]. In theory, our convergence model should be able to make such general predictions. The difference is that we do not know if there is a hidden edge or not. We generate the distances of testing edges as before. Now, instead of classifying the hidden edges in G_r' as positive/negative, we need to classify them as positive or not.

The *general edge prediction*, deciding whether an positive edge exists or not, is

obviously a harder problem than *edge signed prediction*. Social networks are usually sparse, and hence there are a lot more neutral relations than biased relations. If larger distances represent more negativeness (less positiveness), then the distance of a neutral relation should be smaller than a negative one and larger than a positive one. As a consequence, the distribution of neutral edges in terms of distance should concentrate in the middle range distances. We next study this distribution, as a preliminary step to solve the *general edge prediction* problem.

For each dataset, we generate samples based on random source nodes as before, except that we exclude the edges between the k source nodes. Instead of cross validation, we use the entire sub-network for training, and use the $k(k-1)/2$ edges between the k source nodes as testing data, whose signs are available in the original dataset (positive, negative or neutral, i.e. no link). Similarly, the signs of the testing edges should not be influenced strongly by the other edges in the network. Hence, after convergence they should get to the distances corresponding to their true signs. We repeat the experiments 50 times over all three datasets, and collect the distances for only the neutral testing edges.

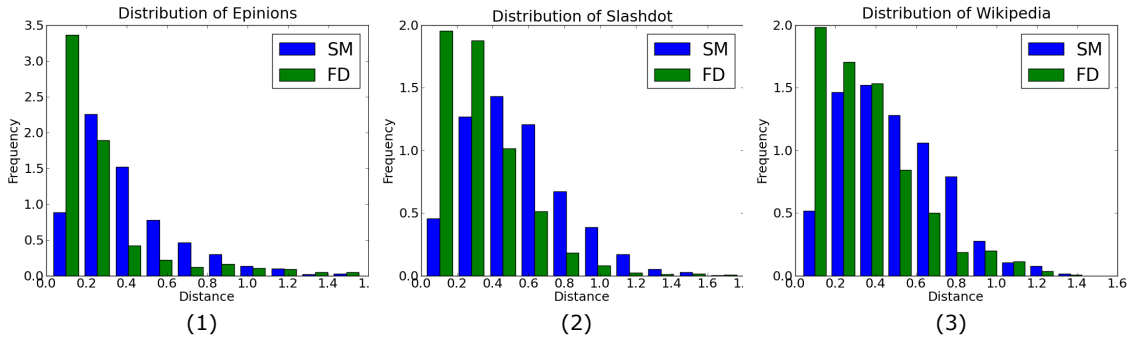


Figure 6.2: The histograms are drawn upon distances of neutral testing edges, generated by SM and FD. (1) (2) (3) correspond to the histograms generated from Epinions, Slashdot and Wikipedia datasets respectively.

As it is shown in Figure 6.2, the distances of neutral testing edges generated by SM do relatively concentrate in the middle-range of values following an almost Gaussian distribution. On the contrast, the majority of neutral testing edges's distances by FD have small values, implying a positive prediction. In fact, based on the dis-

tributions, we can conclude that it is much more probable to classify a missing link as a positive edge in FD than in SM. However, SM provides more flexibility as the distances are distributed over a larger range with an almost Gaussian distribution, allowing us to test different tunable algorithms.

The experimental results show the superiority of our convergence model in predicting signed edges. Moreover, they justify that our general balance theory is a relatively accurate description of the stable state of social networks, and that the convergence model correctly characterizes the dynamics of social network. As a result, our model is a good starting point for developing algorithms for solving the *general edge prediction* problem.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this thesis, we introduced a general model for structural balance theory that can handle relation strengths and generalizes the classical balance theory. We showed that our notion of balance can be mapped to triangular inequality over metric distances and the issue of convergence can be modeled as the metric multidimensional scaling problem for which stress majorization provides exact solutions. We have shown that our theory can be used to effectively solve the edge sign prediction problem and its performance matches and exceeds state of the art for this problem. This is due to the fact that positive and negative edges are mapped to a continuous range of strengths based on the constraints provided by the other nodes. However, in contrast with previous work, our method is aware of global constraints based on balance which results in better results overall. Furthermore, the solutions provided by our method can also be used to solve the harder edge prediction problem.

We are investigating various avenues of future work. Inspired by the work in [25], we have developed an approximation algorithm of stress majorization in the context of social networks. We would like to implement and test how well the approximation is. Furthermore, we are currently testing how the inclusion of relation strength improves performance by considering other actions of the users that imply the existence of a social tie. Our method also has applications to many related problems like clustering and link prediction, which we are currently investigating. We are studying the various properties of our theory in general networks and how it can be extended to an asymmetric interpretation of links. We note that our method not only allows us to make predictions, but also study the characteristics existing networks and compare various characteristics. Example measures would be the ratio between positive and negative distances or between largest and smallest distances. These measures can help us develop new insights into the nature of adversarial relationships in different networks.

REFERENCES

- [1] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York: Cambridge University Press, 2010.
- [2] F. Heider, “Attitudes and cognitive organization,” *J. of Psychology*, pp. 107–12, 1946.
- [3] D. Cartwright and F. Harary, “Structural balance: a generalization of heider’s theory,” *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.
- [4] J. Davis, “Clustering and structural balance in graphs,” *Human Relations*, pp. 181–187, 1967.
- [5] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” in *Proc. of the 19th Int. Conference on World Wide Web*, 2010, pp. 641–650.
- [6] T. DuBois, J. Golbeck, and A. Srinivasan, “Predicting trust and distrust in social networks,” in *IEEE Int. Conference on Social Computing*, MIT, Boston, 2011.
- [7] P. Doreian, “Event sequences as generators of social network evolution,” *Social Networks*, vol. 24, pp. 93–119, 2002.
- [8] M. Granovetter, “The strength of weak ties,” *Amer. J. of Sociology*, vol. 78, pp. 1–22, 1973.
- [9] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of trust and distrust,” in *The 13th Int. Conference on World Wide Web. ACM*, 2004, pp. 403–412.
- [10] E. R. Gansner, Y. Koren, and S. North, “Graph drawing by stress majorization,” *J. Pach (Ed.), Graph Drawing, Lecture Notes in Computer Science*, vol. 3383, pp. 239–250, 2005.

- [11] J. Davis, “Structural balance, mechanical solidarity, and interpersonal relations,” *Amer. J. of Sociology*, vol. 68, pp. 444–462, 1963.
- [12] F. Harary, “On the notion of balance of a signed graph,” *Michigan Math*, vol. 2, no. 2, pp. 143–146, 1953.
- [13] P. Holland and S. Leinhardt, “A dynamic model for social networks,” *Math. Sociology*, vol. 5, pp. 5–20, 1977.
- [14] S. Wasserman, “Analyzing social networks as stochastic processes,” *J. of the Amer. Stat. Assoc.*, vol. 75, pp. 280–294, 1980.
- [15] R. Leenders, “Models for network dynamics: a markovian framework,” *Math. Sociology*, vol. 20, pp. 1–21, 1995.
- [16] T. A. Snijders, “The statistical evaluation of social network dynamics,” *Sociol. Methodology*, pp. 361–395, 2001.
- [17] T. DuBois, J. Golbeck, and A. Srinivasan, “Rigorous probabilistic trust-inference with applications to clustering,” in *Proc. of the 2009 IEEE/WIC/ACM Int. Joint Conference on Web Intell. and Intelligent Agent Technology*, vol. 01. IEEE Computer Society, 2009, pp. 655–658.
- [18] D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks,” in *Int. Conference on Inform. and Knowledge Manage.*, 2003, pp. 556–559.
- [19] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, “Understanding and sharing intentions: The origins of cultural cognition,” *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 675–91; discussion 691–735, 2005.
- [20] S. T. Fiske, A. J. Cuddy, and P. Glick, “Universal dimensions of social cognition: warmth and competence,” *Trends in Cognitive Sciences*, vol. 11, no. 2, pp. 77 – 83, 2007.

- [21] D. Ames, S. Fiske, and A. Todorov, “Impression formation: A focus on others’ intents,” in *The Oxford Handbook of Social Neuroscience*, J. Decety and J. Cacioppo, Eds. New York: Oxford University Press, 2011, pp. 419–433.
- [22] J. D. Leeuw, “Applications of convex analysis to multidimensional scaling,” *Recent Developments in Stat.*, pp. 133–145, 1977.
- [23] J. Kleinberg, “The small-world phenomena: An algorithmic perspective,” in *The 32nd Annu. ACM Symp. on Theory of Computing*. New York: Association of Computing Machinery, 2000, pp. 163–170.
- [24] D. Watts and H. Strogatz, “The small world problem,” *Nature*, pp. 393–440, 1998.
- [25] M. Khoury, Y. Hu, S. Krishnan, and C. Scheidegger, “Drawing large graphs by low-rank stress majorization,” in *EuroVis*, 2012.
- [26] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, “Clustering large graphs via the singular value decomposition,” *Mach. Learning*, vol. 56, no. 9-33, 2004.
- [27] G. Golub and C. Loan, *Matrix Computation*. Baltimore and London: Johns Hopkins University Press, 1996.
- [28] W. Hager, “Updating the inverse of a matrix,” *SIAM Review*, vol. 31, pp. 221–239, June 1989.
- [29] J. Barnes and P. Hut, “A hierarchical $O(n \log n)$ force-calculation algorithm,” *Nature*, vol. 324, no. 4, 1986.

APPENDIX A

Follow Gansner's description, we denote the d-dimensional location matrix as X [10]. Let node i 's location be X_i . The stress function in literature is defined as

$$stress(X) = \sum_{i < j} w_{ij} (||X_i - X_j|| - d_{ij})^2, \quad (\text{A.1})$$

where w_{ij} are weights and d_{ij} denote the ideal distance between i and j . The MDS is an optimization problem to minimize the total stress. Equation A.1 can be decomposed and we have

$$stress(X) = \sum_{i < j} w_{ij} d_{ij}^2 + \sum_{i < j} w_{ij} ||X_i - X_j||^2 - 2 \sum_{i < j} w_{ij} d_{ij} ||X_i - X_j||. \quad (\text{A.2})$$

The first term of A.2 is a constant. The second term, $\sum_{i < j} w_{ij} ||X_i - X_j||^2$, is a quadratic sum, and hence can be written as

$$\sum_{i < j} w_{ij} ||X_i - X_j||^2 = Tr(X^T L^w X),$$

where L^w is a weighted matrix, defined as

$$L_{i,j}^w = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j \end{cases}.$$

The third term, $\sum_{i < j} w_{ij} d_{ij} ||X_i - X_j||$, need extra attention. By Cauchy-Schwartz inequality,

$$||X_i - X_j|| ||Z_i - Z_j|| \geq (X_i - X_j)^T (X_i - Z_j)$$

with the equality when $Z = X$. Hence, we have the following bound of the third term.

$$\sum_{i < j} w_{ij} d_{ij} ||X_i - X_j|| \geq \sum_{i < j} w_{ij} d_{ij} (X_i - X_j)^T (X_i - Z_j) / ||Z_i - Z_j||.$$

Equivalently, the right side of the inequality can be rewritten in the trace form.

$$\sum_{i < j} w_{ij} d_{ij} (X_i - X_j)^T (X_i - Z_j) / \|Z_i - Z_j\| = \text{Tr}(X^T L^Z X),$$

where L^Z is defined as

$$L_{i,j}^Z = \begin{cases} -w_{ij} d_{ij} / \|Z_i - Z_j\| & i \neq j \\ -\sum_{j \neq i} L_{i,j}^Z & i = j \end{cases}.$$

From the above all, we can bound the stress function confidently by function $F^Z(X)$, defined as

$$F^Z(X) = \sum_{i < j} w_{ij} d_{ij}^2 + \text{Tr}(X^T L^w X) - 2\text{Tr}(X^T L^Z Z) \quad (\text{A.3})$$

with equality when $Z = X$. Notice $F^Z(X)$ is of quadratic form, and we can easily find its minima by differentiate it by X and solve

$$L^w X = L^Z Z.$$

This leads to the following iterative optimization process. Given the location matrix at time t , $X(t)$, we want to compute $X(t+1)$ so that $\text{stress}(X(t+1)) < \text{stress}(X(t))$. We take $X(t+1)$ as the minimizer of $F^{X(t)}(X)$ by solving the above equation iteratively. At this point, if $X(t+1) = X(t)$, the process is terminated. Otherwise, we get

$$\text{stress}(X(t+1)) \leq F^{X(t)}(X(t+1)) < F^{X(t)}(X(t)) = \text{stress}(X(t)).$$

To summarize, the majorization process involves solving the following equation iteratively

$$L^w X(t+1) = L^{X(t)} X(t), \quad (\text{A.4})$$

where L^w is constant throughout the process, whereas matrix $L^{X(t)}$ need to be computed at each iteration. While stress majorization is an effective and accurate optimization strategy, it is not feasible in very large networks. Solving (14)

in each iteration requires the computation of a fully dense matrix, which in total requires $O(n^3)$ time and $O(n^2)$ space. For network of node number $> 10^5$, direct implementation of stress majorization becomes impossible.