On Distributed Maximization of Influence in Social Networks

Joakim Guth

Lund University
Department of Automatic Control

MSc Thesis
TFRT-6045
ISSN 0280-5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

Printed in Sweden by Tryckeriet i E-huset
Lund 2017

Abstract

This thesis studies the problem of finding the optimal placement of a directed link in a graph representation of a social network in order to maximize the induced gain of the opinion equilibrium. The model assumes the presence of a set of stubborn nodes and applies a standard DeGroot opinion dynamics model. First we show that an added directed link should point to a stubborn node in order to maximize the impact of the link. The resulting problem reduction then allows for explicit solutions of where the directed link should origin in a few common network topographies such as the line graph and the barbell graph. A formula for the optimal tail placement for general graphs is then presented along with a distributed algorithm. Implementation and simulation are then performed again first on a few common network types and then on a small sub-network of Facebook.

Acknowledgements

I want to thank my supervisors Giacomo Como and Gustav Nilsson for their consistent and extensive interest, support and guidance throughout the whole thesis process. This was absolutely crucial for my work and brought a considerable amount of new knowledge. An especially important contribution from Giacomo was the formulation of the general formula.

I also want to thank my dear friends Martin Vikström Morin and Erik Ottosson with whom I shared work space. They brought the human warmth in my weekdays along with Dennis Leander, Lovisa Nilsson, Bella Gleisner, Elin Törnqvist and many more.

Contents

1

Introduction

During the last decade, the modelling, analysis and synthesis of social systems have gained increasing attention in the area of controls systems research [Friedkin, 2015]. This includes, for example, the study of social learning and opinion dynamics [Acemoglu and Ozdaglar, 2011; Acemoğlu et al., 2013; Como and Fagnani, 2010]. Also the problems of influence maximization in social networks [Kempe et al., 2003; Vassio et al., 2014] that is more closely related to what is treated in this thesis.

This thesis deals with controlling the opinion of people by finding the optimal placement of a link in order to maximize the influence of a given node in a social network. In essence, a social network is composed of a number of persons and co-interacting connections between these or some of these people. A commonly known example of a social network is Facebook, but in reality, any collection of individuals – from two individuals to seven billion– can also be interpreted as a social network.

The main target of this thesis is to find a way to modify a social network in order to sway the opinion of individuals in one way. Consider there being two poles, A and B, of a matter of opinion. This thesis helps a person convinced of for example opinion B to find the optimal person to influence the state of the network towards opinion B.

In Chapter 2 we present general theory where the social network is given a mathematical interpretation. Upon this a variant of the well established DeGroot opinion dynamics model [Degroot, 1974; Friedkin, 2010; Harary, 1959] will be used to model opinion. This variant includes the role of stubborn nodes as introduced in [Acemoğlu et al., 2013; Como and Fagnani, 2016]. Then, in Chapter 3, the problem will be formulated theoretically and the solution will be presented for a few common network topologies. Chapter 4 presents a general formula solving the problem for general networks, followed by a distributed implementation. Chapter 5 presents results from simulations using the algorithm. Finally, in Chapter 6 we draw a few conclusions regarding our results and expectations.

In essence, Chapter 2 is a literature review of well established theory. The important Theorem 4.1 is a contribution from Giacomo Como as well as the use of random walks in approximating the diagonal of K. Remaining developments such as Chapter 3 and Chapter 5 are my own contributions.

1.1 Notation

For the set $\mathcal{V}$, $|\mathcal{V}|$ denotes the number of elements in the set. If a vector x is defined on $\mathbb{R}_+^{\mathcal{V}}$, this is equal to a vector with $|\mathcal{V}|$ elements each defined on $\mathbb{R}_+$. The same definition is used for matrices so that if a matrix A is defined on $\mathbb{R}_+^{\mathcal{V}\times\mathcal{S}}$, then A is a matrix of size $|\mathcal{V}|\times|\mathcal{S}|$ with each element defined on $\mathbb{R}_+$. Also note that $\mathbb{1}$ denotes a column vector with all elements equal to one. This is especially useful when expressing the sum of for example a column vector $x \in \mathbb{R}_+^{\mathcal{V}}$ so that

$$\sum_{i\in\mathcal{V}} x_i = x'\mathbb{1} = \mathbb{1}'x.$$

2

Background

This chapter provides an introduction to relevant graph-theoretical notions and opinion dynamics models in social networks. Initially, the social network will be given a mathematical interpretation in Section 2.1 introducing standard expressions and definitions in the area. The subsequent Section 2.2 introduces how opinion is modelled by using the DeGroot opinion dynamics model. In Section 2.3 this theory is extended with the introduction of stubborn nodes and their impact on the network. Section 2.4 and Section 2.5 show how the opinion dynamics is related to random walks and electrical networks respectively. Finally, Section 2.6 presents some previous work and its connection to the electrical network interpretation.

2.1 Social networks as graphs

A social network can be modelled as a directed weighted graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, where $\mathcal{V}$ is a set of nodes or agents, $\mathcal{E}$ a set of links or edges and W a weight matrix. A simple visualization of a graph is provided in Figure 2.1. In this representation of a graph, nodes are the existing individuals in the network, the links or edges define the direct social connections between these individuals and the weight matrix define the importance of the social connections. The *nodes* in $\mathcal{V}$ are in general labelled as $\{1, \dots, n\}$ where $n = |\mathcal{V}|$ is the number of nodes and also the size of the network.

An *edge* between two nodes $i, j \in \mathcal{V}$ is defined as the pair of nodes $(i, j) \in \mathcal{E}$. The edge is *directed* in such a way that i is influenced by j, this is also known as the edge having its *tail* in i and its *head* in j. This is depicted as

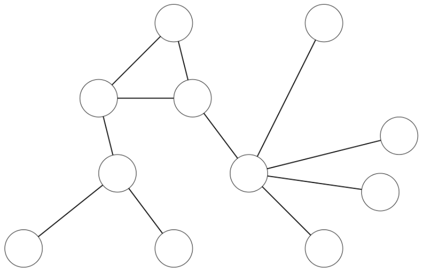


Figure 2.1 A simple network.

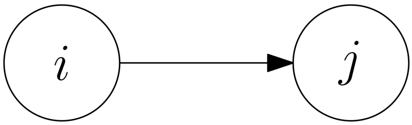


Figure 2.2 Directed link from i to j.

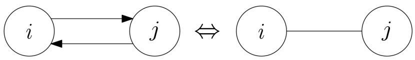


Figure 2.3 Visual interpretation of two directed links.

an arrow in Figure 2.2. Note that when there is also a link from j to i, it is visualized as just a link without any arrow, see Figure 2.3. For a link (i, j), the node j is an *(out)-neighbour* of i and all neighbours to i are said to constitute the *(out)-neighbourhood* of i. If all nodes in the network have each other as neighbours, the network is said to be *complete*, an example can be seen in Figure 2.4. Throughout, it will be assumed that every node has at least one out-neighbour. This causes no essential loss of generality, as we can always add a self-loop (i,i) to every node with no out-neighbours. While, in general, nodes with no out-neighbours other than themselves are known as sinks, they will be referred to as *stubborn nodes* here. These stubborn nodes belong to the subset $\mathcal{S} \subset \mathcal{V}$. All nodes which are not stubborn are called *regular* and belong to the subset $\mathcal{R} = \mathcal{V} \setminus \mathcal{S}$.

The final part of the graph definition is the *weight matrix*, $W \in \mathbb{R}_+^{\mathcal{V}\times\mathcal{V}}$,

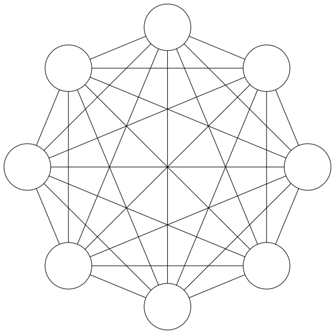


Figure 2.4 A complete graph.

which defines the strength, or the weight, of links. For example, if $(i, j) \in \mathcal{E}$, then $W_{ij} > 0$. The larger this number is, the more influence is exerted on i by j. If there is no edge from a node i to a node j, that is if $(i, j) \notin \mathcal{E}$, then $W_{ij} = 0$. For simplicity, in this thesis we will generally consider unweighted graphs, i.e., graphs such that the weight of every edge is equal to one so that $W \in \{0,1\}^{\mathcal{V}\times\mathcal{V}}$. We shall also refer to a graph as *undirected* if its weight matrix W is symmetric. Otherwise, if W is not symmetric, we shall refer to the graph as *directed*. A matrix related to W, which will be found useful, is the *normalized weight matrix*, P, defined as

$$P = D^{-1}W, \quad D = \text{diag}(w), \quad w = W\mathbb{1}. \tag{2.1}$$

where w_i the out-degree for node i, defined as the aggregate weight of links stemming out of node i. This matrix, P, is stochastic, meaning that the row sum is equal to one and that every entry can be considered a probability.

If there is a path of edges from a node i to node j, then j is said to be *reachable* from i. If all nodes in a network is reachable from each other, then the network is said to be *connected* otherwise it is *disconnected*. An example of a disconnected network can be seen in Figure 2.5. Another example is if there is a stubborn node $j \in \mathcal{S}$ present in the network, no other nodes can be reached from this node and thus the network is disconnected. It will be found useful to define a set of nodes as globally reachable if it is reachable from every node in the network. These statements can be concluded in the following definition.

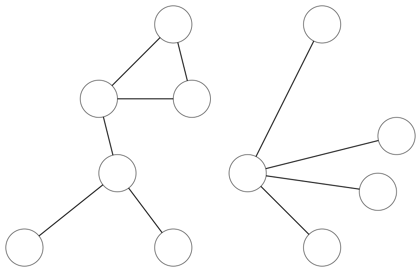


Figure 2.5 A disconnected graph.

DEFINITION 2.1
Given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ and two nodes $i, j \in \mathcal{V}$ then

(i) j is *reachable* from i if there is a path of directed links from i to j.
(ii) the subset $\mathcal{A} \subset \mathcal{V}$ is *reachable* from $i \in \mathcal{V}$ if a node $j \in \mathcal{A}$ is reachable from i.
(iii) the set of nodes $\mathcal{S} \subseteq \mathcal{V}$ is *globally reachable* if it is reachable from every $i \in \mathcal{V}$
(iv) $\mathcal{G}$ is *connected* if $\mathcal{V}$ is globally reachable. □

As the topic is on social influence, it is useful to quantify the importance, or the amount of influence or popularity, a node has in the network. This is commonly known as the centrality of a node and different approaches can be applied to quantify this. If a network is assumed to be connected, then a common measurement is the Bonacich centrality [Bonacich, 1987], $\pi \in \mathbb{R}_{+}^{\mathcal{V}}$, defined as

$$\pi_i = \frac{1}{\lambda} \sum_{j \in \mathcal{V}} W_{ji} \pi_j, \quad i \in \mathcal{V} \tag{2.2}$$

or equivalently expressed in matrix form as

$$\lambda \pi = W' \pi, \quad \pi > 0 \tag{2.3}$$

where $\lambda > 0$ is a proportionality constant. Looking at (2.3), π is an eigenvector of W' and λ the corresponding eigenvalue. According to the Perron-Frobenius theorem for positive matrices [Holst and Ufnarovski, 2014], there is a dominating eigenvalue and the corresponding eigenvector is the only positive eigenvector except for multiples of this. Thus, this rescaling can be done in any way to fit ones purpose. A common way is to choose a rescaling so that $\pi'\mathbb{1} = 1$ and π is then a probability vector. Two factors contribute to a higher centrality of a node in this case, first it is the out-degree and second the centrality of its out-neighbours. An improved version of this is the normalized Bonacich centrality which can be formulated as

$$\pi_i = \frac{1}{\lambda}\sum_j \frac{W_{ji}}{w_j}\pi_j = \frac{1}{\lambda}\sum_j P'\pi_j \tag{2.4}$$

with corresponding matrix form

$$\lambda\pi = P'\pi \tag{2.5}$$

where P is the normalized weight matrix defined in (2.1). As this matrix is stochastic, the largest eigenvalue is equal to one, resulting in

$$\pi = P'\pi\,. \tag{2.6}$$

The normalization is performed because a node with a larger neighbourhood should contribute less to the centrality of a neighbouring node. This is not the case in the unnormalized Bonacich centrality where every neighbour contribute equally to the centrality. Another centrality measurement is the so called PageRank centrality [Brin and Page, 1998] formulated as

$$\pi = (1-\beta)P'\pi + \beta\mu \tag{2.7}$$

where $\beta \in (0,1]$ and $\mu \in \mathbb{R}_+^{\mathcal{V}}$. This was originally developed and used in the Google search engine to provide relevant search results.

2.2 Linear opinion dynamics in social networks

In the previous section, models for the dynamics of influence was intentionally left out but will receive an explanation now with the introduction of the DeGroot opinion dynamics model [Degroot, 1974]. Assume that every node $i \in \mathcal{V}$ holds a value in the interval $x_i \in [0,1]$ which measures their opinion, where 0 and 1 are opposite extremes. All x_i are contained in the opinion (column) vector $x \in \mathbb{R}_+^{\mathcal{V}}$. With DeGroot linear opinion dynamics it is possible to model the social interaction between nodes in a network and how the opinion varies over time. Consider a time depending opinion vector $x(t)$ defined as the opinion of all nodes at time t. Then according to the DeGroot opinion dynamics, nodes update their opinion through the dynamic process

$$x(t+1) = Px(t), \quad x(0) \in \mathbb{R}_+^{\mathcal{V}} \tag{2.8}$$

where P is the normalized weight matrix. Thus, at every time step, every node update their opinion as a linear combination of their neighbours opinions.

As t grows large and approaches infinity it is possible that the opinion distribution will reach an equilibrium so that $x(t+1) = x(t)$. If this equilibrium is denoted x^* then

$$x^* = Px^*. \tag{2.9}$$

Since P is stochastic, and thus have a row sum of one and all of its elements are positive, then again according to the Perron-Frobenius theorem, the only solution to (2.9) is $x^* = c\mathbb{1}$, where $c \in \mathbb{R}_+$ is a positive constant. As one might have observed, all nodes now have the same opinion, in other words a consensus has been reached. It is possible to find a unique c for this problem [Holst and Ufnarovski, 2014] depending on the initial opinions of the nodes. This can be found by using the earlier notion that the normalized Bonacich centrality vector is $\pi = P'\pi$. Multiplication of each side in (2.8) with π' yields

$$\pi' x(t+1) = \pi' Px(t) \tag{2.10}$$

$$\Leftrightarrow \pi' x(t+1) = (P'\pi)' x(t) \tag{2.11}$$

then using the fact from (2.6) that π is stationary results in

$$\pi' x(t+1) = \pi' x(t) \tag{2.12}$$

$$\Leftrightarrow \pi' x(t) = \pi' x(0). \tag{2.13}$$

It is known, since π is a probability vector, that $\pi'\mathbb{1} = \sum_i \pi_i = 1$. Using this fact, and that $x^* = c\mathbb{1}$, in the following statement

$$\begin{aligned}
\pi' x(t) &= \pi' x(0) && (2.14)\\
\Leftrightarrow \pi' x^* &= \pi' x(0) && (2.15)\\
\Leftrightarrow \pi' c\mathbb{1} &= \pi' x(0) && (2.16)\\
\Leftrightarrow c\pi' \mathbb{1} &= \pi' x(0) && (2.17)\\
\Leftrightarrow c &= \pi' x(0) && (2.18)
\end{aligned}$$

returns the solution of c. Assuming that the opinion dynamics, $x(t)$, converges, it will reach the consensus vector $\pi' x(0)\mathbb{1}$ as $t \to \infty$. One more condition is needed for the opinion dynamic to converge. Some networks have an oscillating solution, meaning that there does not exist a unique solution. This is the result of cycles in the network, a cycle being a path from a node back to itself, something that in general are present in a network. For the opinion to converge, it is necessary that the least common divisor of cycle lengths is equal to one, referred to as the graph being aperiodic. The following theorem can now be formulated.

THEOREM 2.1
If the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ is connected and aperiodic where every node follows the DeGroot opinion dynamics

$$x(t+1) = Px(t) \tag{2.19}$$

where $P = D^{-1}W$, $D = \text{diag}(W\mathbb{1})$ and initial condition $x(0) \in \mathbb{R}_+^{\mathcal{V}}$, then the opinion equilibrium is

$$\lim_{t\to\infty} x(t) = x^* = \pi' x(0)\mathbb{1} \tag{2.20}$$

where π is the normalized Bonacich centrality.

Proof [Degroot, 1974] □

2.3 Opinion dynamics with stubborn nodes

What is the result of the existence of stubborn nodes in a network? We model a stubborn node as a node that is not influenced by any other nodes and thus has a constant opinion. Consider the case with only one stubborn node and assume that it is globally reachable. The stubborn node will continuously influence the network until all of the nodes have reached a consensus, which will be equal to the opinion of the stubborn node itself. More interesting and valuable is the case when stubborn nodes of different types are present in the network. In this case, the network can not reach a consensus as it is impossible to change the opinion of the stubborn nodes. The regular nodes though, will rather converge to an opinion influenced by both stubborn nodes. This can be shown through the following. Consider a network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ and let $\mathcal{S}$ be a globally reachable subset of stubborn nodes. Let $u \in \mathbb{R}^{\mathcal{S}}$ be a vector of stubborn nodes' opinions. Refer to $\mathcal{R} = \mathcal{V} \setminus \mathcal{S}$ as the set of regular nodes. Let, $y \in [0,1]^{\mathcal{R}}$, be the opinion vector of all regular nodes. Also, assume that $x(t)$, W and P have been reordered, so that the rows corresponding to the stubborn nodes now is placed last in the matrices. Then, P and $x(t)$ is

$$P = \begin{pmatrix} Q & B \\ C & D \end{pmatrix}, \quad x(t) = \begin{bmatrix} y(t) \\ u \end{bmatrix}. \tag{2.21}$$

Then consider the following opinion dynamics with stubborn nodes

$$x(t+1) = \begin{pmatrix} Q & B \\ 0 & I \end{pmatrix} x(t) \tag{2.22}$$

or as the system of equations

$$\begin{cases} y(t+1) & = Qy(t) + Bu(t) \\ u(t+1) & = u\,. \end{cases} \tag{2.23}$$

Since the opinions of stubborn nodes are constant, $u(t)$ is not changing and can be put as a constant vector u. The matrix $Q \in [0,1]^{\mathcal{R} \times \mathcal{R}}$ is square and substochastic and can be interpreted as the part of P that defines the weighed influence between the regular nodes. The matrix $B \in [0,1]^{\mathcal{R} \times \mathcal{S}}$ can be interpreted as the weighed influence of the stubborn nodes on the regular nodes.

What happens to the opinion distribution as $t \to \infty$? The opinions of the stubborn nodes are, as stated, constant, but what about the regular nodes? Looking at (2.23) and assuming that it is convergent, the equilibrium, y^*, can be found as the following

$$y^* = Qy^* + Bu\,. \tag{2.24}$$

To formulate this converge as a theorem it is useful to present the following lemma about convergence depending the spectral radius of Q.

LEMMA 2.1
Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ be a network where the set $\mathcal{S}$ is globally reachable, then the corresponding block Q, from the normalized weight matrix P as defined in (2.21), has spectral radius

$$\rho(Q) < 1 \tag{2.25}$$

which implies that the inverse of $(I - Q)$ exists, is non-negative and can be computed using the following power series.

$$(I - Q)^{-1} = \sum_{k=0}^{\infty} Q^k. \tag{2.26}$$

Proof [Como and Fagnani, 2016] Since Q is a non-negative matrix, it follows from the Perron Frobenius theory that its spectral radius ρ is an eigenvalue with associated non-negative eigenvector v, i.e., $Qv = \rho v$. Let $\mathcal{J} \subset \mathcal{R}$ be the support of v, that is the non-zero elements for each row in Q. Since P is a row-stochastic matrix with at least one link to a stubborn node and Q is a block of P corresponding to the regular nodes, there is at least one row in Q with row sum of less than one, so that

$$\min_{i \in \mathcal{J}} \sum_{j \in \mathcal{J}} Q_{ij} < 1. \tag{2.27}$$

This implies that

$$\rho \sum_{j \in \mathcal{J}} v_j = \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{R}} Q_{ij} v_i = \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{J}} Q_{ij} v_i < \sum_{i \in \mathcal{J}} v_i \tag{2.28}$$

and hence it must hold that

$$\rho < 1. \tag{2.29}$$

In order to prove the second statement of the lemma, it follows from Gelfand's formula [Holst and Ufnarovski, 2014], that

$$\lim_{k \to \infty} \sqrt[k]{||Q^k||} = \rho < 1 \tag{2.30}$$

and hence the root test can be applied to show that the series

$$\sum_{k=0}^{\infty} Q^k \tag{2.31}$$

converges. Subsequently

$$(I-Q)\sum_{k=0}^{n} Q^k = I - Q^{n+1} \overset{n\to\infty}{\to} I \tag{2.32}$$

which results in

$$\sum_{k=0}^{\infty} Q^k = (I-Q)^{-1} \tag{2.33}$$

□

Using this lemma it is now possible to prove the convergence of the opinion dynamics in a graph with stubborn nodes.

THEOREM 2.2
Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ be a network and $\mathcal{S}$ be a globally reachable set of stubborn nodes. Let $u \in \mathbb{R}^{\mathcal{S}}$ be a vector of stubborn nodes' opinions and let

$$x(t) = \begin{bmatrix} y(t) \\ u \end{bmatrix} \tag{2.34}$$

be the opinion vector of the opinion dynamics (2.22) with stubborn nodes. Then

$$\lim_{t\to\infty} y(t) = (I-Q)^{-1}Bu = y^* \tag{2.35}$$

Proof It follows directly from Lemma 2.1 that

$$\lim_{t\to\infty} y(t) = \lim_{n\to\infty} \sum_{t=0}^{n} Q^t Bu = (I-Q)^{-1}Bu. \tag{2.36}$$

□

2.4 Relation to random walks on networks

The random walk on graphs can be described as the action of randomly walking from node to node in the network. It is then possible to compute the probability to be in a certain node at a certain time. A starting node is chosen for the walk, then at every time step the walk travels along an edge to a randomly selected neighbouring node, where the process continues. Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ with transition probability matrix P, then the random walk on a graph is defined as the stochastic variable $V(t)$ with state space $\mathcal{V}$. Then if a walk starts at node i at time $t = 0$ it randomly chooses a neighbouring node according to the probabilities in P where it will be in $t = 1$. If the walk is in node i at time $t = 0$, then the probability to be in j at time t is

$$\mathrm{P}(V(t) = j | V(0) = i) = (P^t)_{ij}. \tag{2.37}$$

More generally, if the walk starts in node i at $t = 0$, then the probability distribution is a column vector $x(0)$ with all elements zero except $x(0)_i = 1$. Then $x(t)$ will be the probability distribution at time t so that

$$x(t+1) = P' x(t). \tag{2.38}$$

This is the same dynamics as for the normalized Bonacich centrality vector. The relation between random walks and opinion dynamics can be illuminated with the case when two or more stubborn nodes are present so that $\mathcal{S}_1, \mathcal{S}_0 \neq \emptyset$. If a walk ends up in a stubborn node it can not travel from there and is considered to be absorbed. If a walk starts in node i it can then be absorbed by either $\mathcal{S}_0$ or $\mathcal{S}_1$ as t grows. The probability to end up in $\mathcal{S}_1$ instead of $\mathcal{S}_0$ will then coincide with the opinion equilibrium y^*. This can be stated with the following theorem

THEOREM 2.3
Consider the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ where $\mathcal{S}$ is globally reachable and y^* the opinion equilibrium described in (2.24). If $i \in \mathcal{V}$, then the random walk $V(t)$ on $\mathcal{G}$, with $V(0) = i$ and $V(t) = s_1 \in \mathcal{S}_1$ will show the following convergence

$$\lim_{t \to \infty} \mathrm{P}(V(t) = s_1 | V(0) = i) = \left[(I - Q)^{-1} B\right]_{is_1} = y_i^*. \tag{2.39}$$

Proof The probability for a walk starting in i to be in j at time t is described earlier as

$$(P^t)_{ij}. \tag{2.40}$$

As stubborn nodes are present, recall the block partition of P in (2.21) and study the case when $t \to \infty$. Also initially assume that $|\mathcal{S}_1| = 1$.

$$(P^t)_{ij} = \left[\begin{pmatrix} Q & B \\ 0 & I \end{pmatrix}^t\right]_{ij} = \left[\begin{pmatrix} Q^t & Q^{t-1}B + \dots + QB + B \\ 0 & I \end{pmatrix}\right]_{ij} \tag{2.41}$$

$$= \left[\begin{pmatrix} Q^t & (Q^{t-1} + \dots + Q + I)B \\ 0 & I \end{pmatrix}\right]_{ij} \overset{t\to\infty}{\Rightarrow} \left[\begin{pmatrix} 0 & (I-Q)^{-1}B \\ 0 & I \end{pmatrix}\right]_{ij}. \tag{2.42}$$

It can then be observed that the random walk will be absorbed in the stubborn nodes as $t \to \infty$. Then

$$\lim_{t\to\infty} \mathrm{P}(V(t) = s_1 | V(0) = i) = \left[(I-Q)^{-1}B\right]_{is_1}. \tag{2.43}$$

If $\mathcal{S}_1$, the probability to reach any node in $j \in \mathcal{S}_1$ from i is

$$\sum_{j\in\mathcal{S}_1} \left[(I-Q)^{-1}B\right]_{ij} \tag{2.44}$$

coincidentally, this is equal to

$$\left[(I-Q)^{-1}B\right]_{ij} u = y^* \tag{2.45}$$

where u is the opinion of the stubborn nodes. □

Later on, the sum of the probability for a walk to return to its origin node will be useful. The following theorem shows an alternative way for that expression.

THEOREM 2.4
Consider the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ where $\mathcal{S}$ is globally reachable. If a random walk $V(t)$ on $\mathcal{G}$ with $V(0) = i$ has the probability p_i to ever return to i then

$$\sum_{k=0}^{\infty} (Q^k)_{ii} = \frac{1}{1-p_i}. \tag{2.46}$$

Proof [Gravner, 2014] Begin by study the following equality

$$\sum_{k=0}^{\infty}(Q^k)_{ii} = \sum_{k=0}^{\infty}\mathrm{P}(V(k)=i|V(0)=i) = \mathrm{E}\left(\sum_{k=0}^{\infty}(I_k|V(0)=i)\right) \tag{2.47}$$

where $I_k = I_{V(k)=i}, k = 0,1,2\ldots$ is the indicator function. The last equality in (2.47) shows that the sum of the return probability is equal to the expected number of returns. The expected number of returns has a geometric distribution since the probability for n returns is

$$p_i^{n-1}(1-p_i). \tag{2.48}$$

Then the expected number of returns is

$$\mathrm{E}\left(\sum_{k=0}^{\infty}(I_k|V(0)=i)\right) = \frac{1}{1-p_i}. \tag{2.49}$$

Thus,

$$\sum_{k=0}^{\infty}(Q^k)_{ii} = \frac{1}{1-p_i}. \tag{2.50}$$

□

2.5 The electrical network interpretation

In this section we will present another interpretation of the opinion dynamics with stubborn nodes, namely as an electrical network. Although the electrical network interpretation will not be utilized in the development of the results in this thesis, it gives insights about how similar previous work is done.

In a network with two poles of an opinion such as the two types of stubborn nodes, it is possible to translate it into an electrical network. The opinion of every node can be considered to be voltages in the network and all link weights conductances. The result of this interpretation is that it is possible to use laws and operations normally used on electrical networks such as the laws of Ohm and Kirchoff.

If the conductance be defined as $C_{ij} = W_{ij}$ in the matrix C, then if every edge is numbered $1, 2, \dots, |\mathcal{E}|$ the conductances can be placed on the diagonal in the diagonal matrix $D_C \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$. The voltages $V \in \mathbb{R}_+^{\mathcal{V}}$ are equal to the opinion of every node so that $V = x$. Also assume an incidence matrix $B \in \{0, 1, -1\}^{\mathcal{E}\times\mathcal{V}}$ so that for every edge $e_k = (i, j) \in \mathcal{E}$, $B_{ki} = 1$ and $B_{kj} = -1$. In other words, every row is an edge with entry 1 for the tail placement and -1 for the head placement. Let $\Phi \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the current flow along each edge and $\eta \in \mathbb{R}^{\mathcal{V}}$ the external current input or output. Then according to Ohms law

$$\Phi = D_C B V \tag{2.51}$$

and Kirchoffs law

$$B'\Phi = \eta. \tag{2.52}$$

Noting that resistance is the inverse of conductance, it is possible to use something called effective resistance, so that if two nodes i and j are in two different parts of the network, the paths between them can be collapsed to one with a certain effective resistance using serial and parallel laws. Since nodes with the same voltage has zero resistance towards each other, it is possible to rewrite all these same-voltage nodes as one, called gluing.

2.6 Previous work

Among the previous work done in the field of opinion dynamics, the work by DeGroot is especially important, who studied conditions for social consensus in groups of people using mathematical models. The method used in this thesis to model social interactions in networks in is in fact called DeGroot opinion dynamics.

The main problem in this thesis deals with modifying the network or the social group in a way that can be interpreted as introducing a person in a community and expose a person to that new persons opinions. In more detail finding the optimal person to be exposed so that the new persons opinion spreads the most. Previous work in this area deals with a variation of this problem and rather tries to find the best person in the network to radicalize in order to reach the same goal, analogous to transform a node from regular to stubborn. Even if the problem is similar, the simplest example will later show that the solution differs. For example in [Vassio et al., 2014], they use the electrical network interpretation. They derived an algorithm using something called message passing with extensive use of effective resistance. The benefit of doing this is that the impact of turning a node into a stubborn can be computed by just knowing the impact of its neighbours. The resulting algorithm solves the problem of finding the optimal person perfectly on certain network types and, after modifications are done, at least gives an approximation on general networks. This algorithm is later proved to be convergent [Rossi and Frasca, 2016], but it does not converge to the right solution for all types of networks. However, the solution might still be sufficient enough to find the optimal placement in relation to other options.

3

Influence maximization problem

For various reasons, there could be a desire to affect the opinion distribution of a network, for example, to favour one opinion or to reduce polarization in the network. This could potentially be achieved by modifications in the network such as an artificial addition or removal of a connection between two or more individuals, in theory done by adding or removing a link. Other changes could include increasing or decreasing the exposure to already present connections by increasing or decreasing the current link weight. In this chapter, the first case will be studied – how the addition of a directed connection in the network can benefit in the objective to maximize the total opinion distribution. That is, to find the optimal position of this link to skew the opinion of nodes towards one side. A theoretical representation of this will be presented in Section 3.1. The subsequent Section 3.2 will show that the optimal placement of the head, of such a directed link, is to a stubborn node of the desired opinion and thus simplify the problem. The last Section 3.3 finds the solution for a few common network topologies, such as a line and a barbell.

3.1 Theoretical formulation of the influence maximization problem

The problem formulation can be given a theoretical representation through the following. Consider a nominal network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ where a directed link then is added from node i to node j to create the modified counter part $\tilde{\mathcal{G}} = (\mathcal{V}, \tilde{\mathcal{E}}, \tilde{W})$. The action of adding a does not change the node set, why the node sets are equal. The set of edges and the weight matrix will be modified in the following way

$$\tilde{\mathcal{E}} = \mathcal{E} \cup \{(i,j)\} \tag{3.1}$$

$$\tilde{W} = W + \delta^{(i,j)} \tag{3.2}$$

and the resulting changes to the normalized weight matrix will change the blocks $\tilde{Q}$ and $\tilde{B}$ as

$$\tilde{Q}(i,j) = \delta^{(i,\cdot)} \frac{w_i}{w_i+1} Q + \delta^{(i,j)} \frac{1}{w_i+1} \tag{3.3}$$

$$\tilde{B}(i,j) = \delta^{(i,\cdot)} \frac{w_i}{w_i+1} B + \delta^{(i,j)} \frac{1}{w_i+1}\,. \tag{3.4}$$

Assume the stubborn node set $\tilde{\mathcal{S}} \subset \mathcal{V}$ with two different kind of stubborn nodes $\tilde{\mathcal{S}}_0 \subset \tilde{\mathcal{S}}$ and $\tilde{\mathcal{S}}_1 \subset \tilde{\mathcal{S}}$. Then the objective is to modify the network in such a way that the opinion equilibrium changes in favour of the opinion of $\tilde{\mathcal{S}}_1$. For modelling purposes however, assume the earlier defined stubborn nodes with opinion 0 and 1, so that $\tilde{\mathcal{S}}_0 = \mathcal{S}_0$ and $\tilde{\mathcal{S}}_1 = \mathcal{S}_1$. The maximization problem can then be formulated as

$$\max_{i\in\mathcal{R}, j\in\mathcal{V}} \sum_{k\in\mathcal{R}} \tilde{y}^*_k = \max_{i\in\mathcal{R}, j\in\mathcal{V}} \sum_{k\in\mathcal{R}} \left[(I - \tilde{Q}(i,j))^{-1}\tilde{B}(i,j)u\right]_k. \tag{3.5}$$

where $\tilde{y}^*$ is the opinion equilibrium in $\tilde{\mathcal{G}}$. Thus it is a two dimensional optimization problem since it is necessary to locate both an optimal position for the head, j, and the tail, i, of the directed link. The next section will show that it is possible to reduce the problem to one variable.

3.2 Where should the additional link point to?

Knowing a priori where to connect the head would contribute to reducing the complexity of the problem. A logical assumption would be to place the head in a node with as strong opinion as possible. In this case would be a stubborn node $s_1 \in \mathcal{S}_1$ with the opinion $x_{s_1} = 1$. The following lemma proves that the

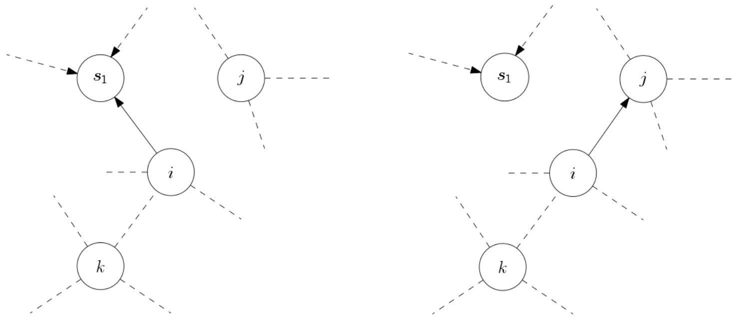


Figure 3.1 The two cases where to connect the head, in s_1 or in j.

sum of the opinion distribution will be higher if the head is placed in $\mathcal{S}_1$ than if it is not. Figure 3.1 provides a visualization of these choices, where the networks are equal except for the two different link placements.

LEMMA 3.1
Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ be a network and let $\mathcal{S} \subset \mathcal{V}$ be a globally reachable set. Let $s_1 \in \mathcal{S}_1$ and $u \in \mathbb{R}^{\mathcal{S}}$ be such that $u_s = 1$ for $s \in \mathcal{S}_1$ and $u_s = 0$ for all $s \in \mathcal{S}_0$. Let $\tilde{\mathcal{G}} = (\mathcal{V}, \tilde{\mathcal{E}}, \tilde{W})$ be such that $\tilde{W} = W - \delta^{(i,j)} + \delta^{(i,s_1)}$, $i, j \in \mathcal{V} \setminus \mathcal{S}$. Let $x^* \in \mathbb{R}_+^{\mathcal{V}}$ and $\tilde{x}^* \in \mathbb{R}_+^{\mathcal{V}}$ be the opinion equilibrium for $\mathcal{G}$ and $\tilde{\mathcal{G}}$ respectively corresponding to the stubborn node opinion vector u. Then the following is true

$$\tilde{x}_k^* \geq x_k^*, \quad k \in \mathcal{V}. \tag{3.6}$$

Proof Consider a Markov coupling of two random walks $V(t)$ and $\tilde{V}(t)$ on $\mathcal{V}$ with transition probability matrix P and $\tilde{P}$, respectively, defined as the following. When $V(t) \neq \tilde{V}(t)$ they move independently with transition probabilities P and $\tilde{P}$, respectively and when $V(t) = \tilde{V}(t) = h \neq i$, $V(t)$ and $\tilde{V}(t)$ move together to a new state $V(t+1) = \tilde{V}(t+1) = k$ chosen with probability $P_{hk} = \tilde{P}_{hk}$. When $V(t) = \tilde{V}(t) = i$, then

$$\mathbb{P}(V(t+1) = \tilde{V}(t+1) = k | V(t) = \tilde{V}(t) = i) = P_{ik} = \tilde{P}_{ik} \qquad k \neq j$$

$$\mathbb{P}(V(t+1) = j, \tilde{V}(t+1) = s_1 | V(t) = \tilde{V}(t) = i) = P_{ij} = \tilde{P}_{is_1}.$$

It is then clear that, if started from the same node k, it can never occur that $V(t)$ is absorbed in s_1 when $\tilde{V}(t)$ is not, so that

$$\begin{aligned}\mathbb{P}(V(t)=s_1|V(0)=k) &= \mathbb{P}(V(t)=s_1|V(0)=k,\tilde{V}(0)=k) && (3.7)\\ &\leq \mathbb{P}(\tilde{V}(t)=s_1|V(0)=k,\tilde{V}(0)=k) && (3.8)\\ &= \mathbb{P}(\tilde{V}(t)=s_1|\tilde{V}(0)=k) && (3.9)\end{aligned}$$

for all $t \geq 0$. It then follows from Theorem 2.4 that

$$x_k = \lim_{t\to+\infty} \mathbb{P}(V(t)=s_1|V(0)=k) \leq \lim_{t\to+\infty} \mathbb{P}(\tilde{V}(t)=s_1|\tilde{V}(0)=k) = \tilde{x}_k\,. \tag{3.10}$$

□

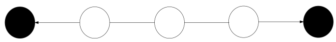
The lemma shows that it is indeed more beneficial to put the head of a directed link in a stubborn node in order to maximize the total opinion distribution. This fact simplifies the previously stated problem of finding i and j in $\mathcal{G}$. Thus, the head j will always be assumed to be in the set $\mathcal{S}_1$ and it is only necessary to find the placement of the tail i.

3.3 Where should the additional link stem from?

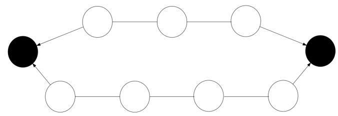
As it is now known that the head of an added directed link, should be placed in a stubborn node belonging to $\mathcal{S}_1$, a logical extension would be to explicitly derive the tail placement for a few simpler network models. This is an effective way to build intuition and understanding on the path to a possible general solution. Initially, the line graph, Figure 3.2(a), will be studied for any choice of n. Then these results will be used to find the placement when two line graphs of different size are present, Figure 3.2(b). Finally the barbell graph, Figure 3.2(c), will be studied. During these derivations it will also be assumed that every network contain one stubborn node of each extreme opinion so that $|\mathcal{S}_0| = |\mathcal{S}_1| = 1$.

Line graph

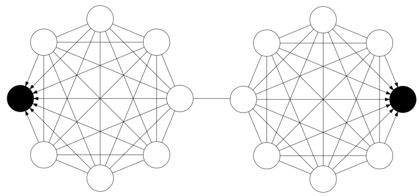
The problem of finding the optimal link placement in a line graph is depicted in Figure 3.3. The following proposition shows and proves the explicit solution on this graph. Some assumptions in this proof uses Matlab simulation



(a) Line graph with $n = 5$ and two stubborn nodes.



(b) Two line graphs and two stubborn nodes.



(c) Barbell graph with $n = 16$ and two stubborn nodes.

Figure 3.2

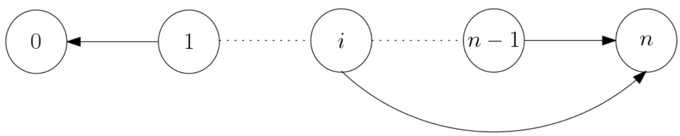


Figure 3.3 A line graph with $n+1$ nodes and a directed link from node i to n.

results presented in Figure 3.4. The figure depicts an approximation of the solution computed with the built in Gaussian elimination based linear systems solver. The curve $\sqrt{n}-1$ is added as a comparison.

PROPOSITION 3.1
Consider the graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, where $\mathcal{V} = \{0, 1, 2, ..., n-1, n\}$ and $\mathcal{E} = (1,0) \cup \{1,2\} \cup \ldots \{n-2, n-1\} \cup (n-1, n)$ and also where $\mathcal{S}_0 = 0$ and $\mathcal{S}_1 = n$. Assume that $n > 7$. If a directed link (k, n) is added, then (3.5) is fulfilled for $k \in (\sqrt{n} - \frac{5}{2}, \sqrt{n})$.

Proof Let x be the opinion vector of the line network with $n+1$ nodes and two stubborn nodes with opinion $x_0 = 0$ and $x_1 = 1$. A link is added from node $k \in \mathcal{V} \setminus \mathcal{S}$ to node $n \in \mathcal{S}_1$. Let $x^{(k)}$ be the opinion of the network with the added

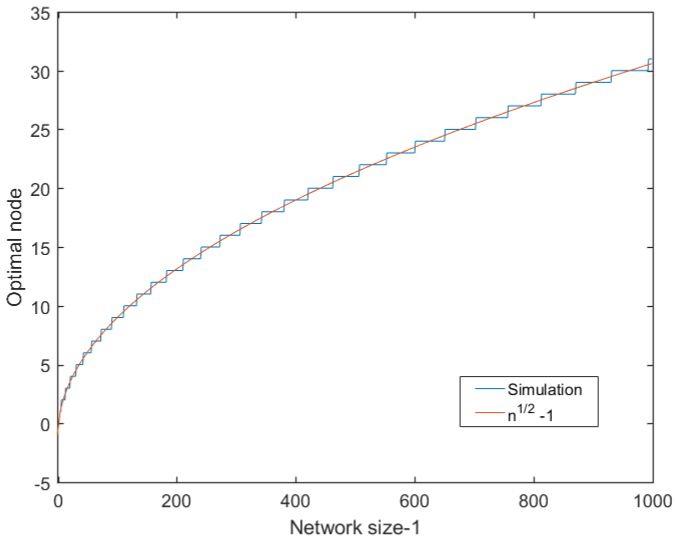


Figure 3.4 Optimal tail placement in line graphs of increasing size.

link. The action of adding a link results in the following average opinion gain

$$\gamma_k = \frac{1}{n+1}\sum_{0\le j\le n} x_j^{(k)} - \frac{1}{n+1}\sum_{0\le j\le n} x_j. \tag{3.11}$$

Since the opinion changes linearly from node 0 to node n, the opinion of the unmodified network is

$$x_j = \frac{j}{n}, \qquad 0 \le j \le n. \tag{3.12}$$

Using the same linear assumption about the opinion after the link is added gives the following result

$$x_j^{(k)} = \frac{j}{k} x_k^{(k)} \quad 0 \le j \le k \tag{3.13}$$

$$x_j^{(k)} = x_k^{(k)} + \frac{j-k}{n-k}(1 - x_k^{(k)}) \quad k \le j \le n \tag{3.14}$$

with the opinion of the node with the added link as

$$x_k^{(k)} = \frac{1}{3}\left(x_{k+1}^{(k)} + x_{k-1}^{(k)} + 1\right). \tag{3.15}$$

The combination of the above equations (3.14), and (3.15), results in

$$x_k^{(k)} = \frac{k(1+n-k)}{n+kn-k^2}. \tag{3.16}$$

The average opinion gain is then, with the insertion of $x^{(k)}$, equal to

$$\begin{aligned}
\gamma_k &= \frac{1}{n+1}\sum_{j=0}^{n} x_j^{(k)} - \frac{1}{n+1}\sum_{j=0}^{n} x_j & (3.17)\\
&= \frac{1}{n+1}\left(\sum_{j=0}^{k}\frac{j}{k}x_k^{(k)} + \sum_{j=k+1}^{n}\left(\frac{1-x_k^{(k)}}{n-k}(j-k)+x_k\right) - \sum_{j=0}^{n}\frac{j}{n}\right) & (3.18)\\
&= \frac{x_k^{(k)}(k+1)}{2(n+1)} + \frac{(1-x_k^{(k)})(n-k+1)}{2(n+1)} + \frac{x_k^{(k)}(n-k)}{n+1} - \frac{1}{2} & (3.19)\\
&= \frac{nx_k^{(k)}-k}{2(n+1)} & (3.20)\\
&= \frac{k(n-k)^2}{2(n+1)(n+nk-k^2)} & (3.21)\\
&= \frac{n^2}{2(n+1)}g(k/n). & (3.22)
\end{aligned}$$

where

$$g(x) = \frac{x(1-x)^2}{1+nx-nx^2}. \tag{3.23}$$

Hence,

$$\max_{k=0,1,\ldots,n}\gamma_k = \frac{n^2}{2(n+1)}\max_{x=0,\frac{1}{n},\frac{2}{n},\ldots,1} g(x). \tag{3.24}$$

We are then interested in relaxed problem of finding the maximum of $g(x)$ on the interval $[0,1]$. Towards this goal, first note that differentiation of yields

$$g'(x) = \frac{(1-x)h(x)}{(nx^2-nx-1)^2}\,, \qquad h(x) = nx^2(x-1)-3x+1\,, \tag{3.25}$$

Since the term $(1-x)/(nx^2-nx-1)^2$ is positive for $x \in (0,1)$, we are interested in the sign of $h(x)$. Note that, $h(x)$ is a third-order polynomial and, for $n \geq 2$, it satisfies

$$h(-1) = 2-2n < 0\,, \qquad h(0) = 1 > 0\,, \tag{3.26}$$
$$h(1) = -2 < 0\,, \qquad h(2) = 4n-5 > 0\,. \tag{3.27}$$

Hence, $h(x)$ has exactly one zero in each of the intervals $(-1,0)$, $(0,1)$, and $(1,2)$, respectively. We will now show that

$$h\left(\frac{\sqrt{n}-\frac{3}{2}}{n}\right) > 0 \qquad h\left(\frac{\sqrt{n}-1}{n}\right) < 0. \tag{3.28}$$

so that the only zero of $h(x)$ in the interval $(0,1)$ actually belongs to the subinterval $(\frac{\sqrt{n}-\frac{3}{2}}{n}, \frac{\sqrt{n}-1}{n})$. Towards this end, observe that

$$\begin{aligned} h\left(\frac{\sqrt{n}-a}{n}\right) &= \frac{1}{n^2}\left(\sqrt{n}-a\right)^3 - \frac{1}{n}\left(\sqrt{n}-a\right)^2 - \frac{3}{n}\left(\sqrt{n}-a\right)+1 && (3.29)\\ &= \frac{1}{n^2}\left(2n^{3/2}(a-1)-na^2+3a^2\sqrt{n}-a^3\right). && (3.30) \end{aligned}$$

Then,

$$h\left(\frac{\sqrt{n}-1}{n}\right) = -n+3\sqrt{n}-1 < 0\,, \qquad n \geq 7\,, \tag{3.31}$$

$$h\left(\frac{\sqrt{n}-3/2}{n}\right) = \frac{n^{3/2}}{2} - \frac{9}{4}n + \frac{27}{4}\sqrt{n} - \frac{27}{8} > 0\,, \qquad n \geq 7\,. \tag{3.32}$$

Hence, the only zero of $h(x)$ in the interval $(0,1)$ actually belongs to the subinterval $(\frac{\sqrt{n}-\frac{3}{2}}{n}, \frac{\sqrt{n}-1}{n})$ so that

$$\operatorname*{argmax}_{x\in[0,1]} g(x) \in \left(\frac{\sqrt{n}-\frac{3}{2}}{n}, \frac{\sqrt{n}-1}{n}\right). \tag{3.33}$$

This in turn implies that

$$\underset{k=0,1,\ldots,n}{\operatorname{argmax}}\, g(x) \in \left(\sqrt{n} - \frac{5}{2}, \sqrt{n}\right). \tag{3.34}$$

□

Multiple lines

There might be cases where multiple line graphs are present. If a directed link is added to a stubborn node in order to maximize the sum of the opinion equilibrium, where should the tail then be placed? The following proposition proves that it is to be placed in the longest one, that is, the one with the most nodes. This is done by showing that the total opinion distribution increase when the network size increases.

PROPOSITION 3.2
Consider a line graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ of size n with two stubborn nodes of different opinion placed in the end points as described in Proposition 3.1 and a directed link with the head in $\mathcal{S}_1$ and the tail in node $(\sqrt{n}-1)$. If n increases, then the sum of the opinion equilibrium x^* increases for $n > 7$.

Proof Showing that the sum of the opinion equilibrium is larger when the size of the line graph increases is equivalent to showing that the derivative of (3.23) for $n > 7$ is strictly positive. Starting from (3.23) and inserting $x = \frac{k}{n}$ results in

$$g(k) = \frac{k(k^2 - 2nk + n^2)}{n + nk - k^2} \tag{3.35}$$

and assume $k = \sqrt{n} - 1$ as this is close to the middle of the optimal interval in Proposition 3.1. Also substitute $\sqrt{n} = s$, then

$$g(s) = \frac{s^5 - 3s^4 + 5s^3 - 5s^2 + 3s - 1}{s^3 - 2s^2 + 2s}. \tag{3.36}$$

Differentiation yields the following

$$g'(s) = \frac{\left(s^2 - s + 1\right)\left(2s^5 - 7s^4 + 11s^3 - 5s^2 - 2s + 2\right)}{s^2\left(s^2 - 2s + 2\right)^2} \tag{3.37}$$

where the sign depending on s now can be studied. First term in the numerator and the whole denominator is greater than zero for $s \geq 1$. Left is to prove that the right part of the numerator also is greater or equal to zero. This can be shown through the following:

$$
\begin{aligned}
&\left(2s^5 - 7s^4 + 11s^3 - 5s^2 - 2s + 2\right) && (3.38)\\
&= 2s^3\left(s^2 - \frac{7}{2}\right) + 11s^3 - 5s^2 - 2s + 2 && (3.39)\\
&= 2s^3\left((s - \frac{7}{4})^2 - \frac{49}{16}\right) + 11s^3 - 5s^2 - 2s + 2 && (3.40)\\
&= 2s^3(s - \frac{7}{4})^2 + \frac{39}{8}s^3 - 5s^2 - 2s + 2 && (3.41)\\
&= 2s^3(s - \frac{7}{4})^2 + \frac{39}{8}s\left(s^2 - \frac{40}{39}s - \frac{16}{39}\right) + 2 && (3.42)
\end{aligned}
$$

which is greater than zero whenever

$$
s^2 - \frac{40}{39}s - \frac{16}{39} > 0 \Leftrightarrow s > \frac{20}{39} + \sqrt{\frac{20^2}{39^2} + \frac{16}{39}} = \frac{4}{3} \Leftrightarrow n > \frac{16}{9} \approx 2\,. \quad (3.43)
$$

□

It has now been shown that the total opinion gain is larger when the tail of the additional link is connected to a longer line graph.

Barbell graph

The barbell graph consists of two complete graphs, each containing $\frac{n}{2}$ nodes, where one node in one side is linked to a node in the other side. Then, to align this with the model, one node in each side of the barbell is turned into a stubborn node, see Figure 3.5. One, s_0, with opinion 0, placed in side A, and one, s_1, with opinion 1, placed in side B. There are now four different kind of nodes where the tail of the directed link could be placed:

- a - regular node in side A
- a_0 - regular node in side A but connected to node b_0 in side B
- b - regular node in side B

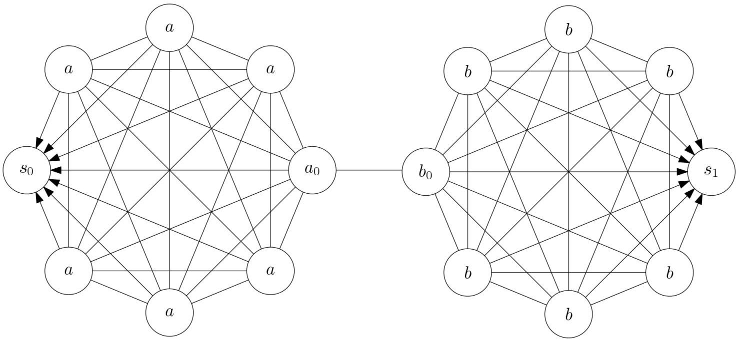


Figure 3.5 A barbell graph with two stubborn nodes of opposing opinion. Left part is side A and right part is side B.

- b_0- regular node in side B but connected to node a_0 in side A.

The following proposition proves that the optimal placement of the tail in this case is from node a_0 to maximize the sum of the opinion equilibrium.

PROPOSITION 3.3
Consider two complete graphs $\mathcal{G}_A = (\mathcal{V}_A, \mathcal{E}_A, W_A)$ and $\mathcal{G}_B = (\mathcal{V}_B, \mathcal{E}_B, W_B)$, where $|\mathcal{V}_A| = |\mathcal{V}_B| = n/2, n \geq 6$, modified so that $s_0 \subset \mathcal{V}_A$ and $\mathcal{S}_1 \subset \mathcal{V}_B$, $|\mathcal{S}_0| = |\mathcal{S}_1| = 1$. Resulting in no out-going links from $\mathcal{S}_0$ or $\mathcal{S}_1$. Then let the nodes $a_0 \in \mathcal{V}_A \setminus \mathcal{S}_0$ and $b_0 \in \mathcal{V}_B \setminus \mathcal{S}_1$ be connected through the edges (a_0, b_0) and (b_0, a_0). Then if a directed link is added so that $(i, s_1), i \in \mathcal{V}_A \cup \mathcal{V}_B \setminus \mathcal{S}, s_1 \in \mathcal{S}_1$, the choice of i that maximizes the sum of the opinion equilibrium is a_0.

Proof If the node, from where a directed link is added, is denoted, i, then the total opinion can be calculated for each of the four cases.

Case 1: Tail of stubborn node connected to node of type a First a system of equations is set up with the expression of the opinion of every present kind of node. Note that $x_{s_0} = 0$ and $x_{s_1} = 1$, e.g. the opinion of the stubborn nodes. Also x_i is the opinion of the node i, or the node of type a where the tail of the directed link is added.

$$\begin{cases} x_i &= \frac{1}{n/2}\left((\frac{n}{2}-3)x_a + x_{a_0} + x_{s_0} + x_{s_1}\right) \\ x_a &= \frac{1}{n/2-1}\left((\frac{n}{2}-4)x_a + x_{s_0} + x_{a_0} + x_i\right) \\ x_{a_0} &= \frac{1}{n/2}\left((\frac{n}{2}-3)x_a + x_{s_0} + x_{b_0} + x_i\right) \\ x_b &= \frac{1}{n/2-1}\left((\frac{n}{2}-3)x_b + x_{s_1} + x_{b_0}\right) \\ x_{b_0} &= \frac{1}{n/2}\left((\frac{n}{2}-2)x_b + x_{s_1} + x_{a_0}\right) \end{cases} \Leftrightarrow \begin{cases} x_i &= \frac{6n^2+24n-16}{n^3+12n^2+24n-16} \\ x_a &= \frac{4n^2+12n-16}{n^3+12n^2+24n-16} \\ x_{a_0} &= 2\,\frac{3n^2+6n-16}{n^3+12n^2+24n-16} \\ x_b &= \frac{n^3+10n^2+20n-24}{n^3+12n^2+24n-16} \\ x_{b_0} &= \frac{n^3+8n^2+16n-32}{n^3+12n^2+24n-16} \end{cases} \tag{3.44}$$

Then the total opinion is given by the sum

$$x_{tot}^{(a)} = x_i + \left(\frac{n}{2}-3\right)x_a + x_{a_0} + \left(\frac{n}{2}-2\right)x_b + x_{b_0} + x_{s_0} + x_{s_1} \tag{3.45}$$

$$= \frac{n\left(n^3+14n^2+32n-40\right)}{2n^3+24n^2+48n-32} \tag{3.46}$$

In the same way, remaining cases can be solved

Case 2: Tail of stubborn node connected to node of type a_0 Expressions

$$\begin{cases} x_a &= \frac{1}{n/2-1}\left((\frac{n}{2}-3)x_a + x_{s_0} + x_{a_0}\right) \\ x_i = x_{a_0} &= \frac{1}{n/2+1}\left((\frac{n}{2}-2)x_a + x_{s_0} + x_{b_0} + x_{s_1}\right) \\ x_b &= \frac{1}{n/2-1}\left((\frac{n}{2}-3)x_b + x_{s_1} + x_{b_0}\right) \\ x_{b_0} &= \frac{1}{n/2}\left((\frac{n}{2}-2)x_b + x_{s_1} + x_{a_0}\right) \end{cases} \Leftrightarrow \begin{cases} x_a &= 4\,\frac{n+2}{n^2+12n+16} \\ x_i = x_{a_0} &= 8\,\frac{n+2}{n^2+12n+16} \\ x_b &= \frac{n^2+10n+16}{n^2+12n+16} \\ x_{b_0} &= \frac{(n+4)^2}{n^2+12n+16} \end{cases} \tag{3.47}$$

Total opinion

$$x_{tot}^{(a_0)} = \frac{n^3+18n^2+84n+96}{2n^2+24n+32} \tag{3.48}$$

Case 3: Tail of stubborn node connected to node of type b In this case x_i is the opinion of the node i, or the node of type b where the tail of the directed link is added. Expressions

$$\begin{cases} x_a &= \frac{1}{\frac{n}{2}}\left(\left(\frac{n}{2}-3\right)x_a + x_{s_0} + x_{a_0}\right) \\ x_{a_0} &= \frac{2}{n}\left(\left(\frac{n}{2}-2\right)x_{a_0} + x_{s_0} + x_{b_0}\right) \\ x_b &= \frac{1}{\frac{n}{2}-1}\left(\left(\frac{n}{2}-4\right)x_b + x_i + x_{b_0} + x_{s_1}\right) \\ x_{b_0} &= \frac{2}{n}\left(\left(\frac{n}{2}-3\right)x_b + x_{a_0} + x_i + x_{s_1}\right) \\ x_i &= \frac{2}{n}\left(\left(\frac{n}{2}-3\right)x_b + x_{b_0} + 2x_{s_1}\right) \end{cases} \Leftrightarrow \begin{cases} x_a &= \frac{2n+8}{n^2+12n+28} \\ x_{a_0} &= \frac{4n+16}{n^2+12n+28} \\ x_b &= \frac{n^2+10n+24}{n^2+12n+28} \\ x_{b_0} &= \frac{(n+4)^2}{n^2+12n+28} \\ x_i &= \frac{n^2+10n+28}{n^2+12n+28} \end{cases} \tag{3.49}$$

Total opinion

$$x_{tot}^{(b)} = \frac{n(n+8)(n+4)}{2n^2+24n+56} \tag{3.50}$$

Case 4: Tail of stubborn node connected to node of type b_0 Expressions

$$\begin{cases} x_a &= \frac{1}{\frac{n}{2}}\left(\left(\frac{n}{2}-3\right)x_a + x_{s_0} + x_{a_0}\right) \\ x_{a_0} &= \frac{2}{n}\left(\left(\frac{n}{2}-2\right)x_{a_0} + x_{s_0} + x_{b_0}\right) \\ x_b &= \frac{1}{\frac{n}{2}-1}\left(\left(\frac{n}{2}-3\right)x_b + x_{b_0} + x_{s_1}\right) \\ x_i = x_{b_0} &= \frac{1}{\frac{n}{2}+1}\left(\left(\frac{n}{2}-2\right)x_b + x_{a_0} + 2x_{s_1}\right) \end{cases} \Leftrightarrow \begin{cases} x_a &= \frac{2n+8}{n^2+12n+16} \\ x_{a_0} &= \frac{4n+16}{n^2+12n+16} \\ x_b &= \frac{n^2+10n+16}{n^2+12n+16} \\ x_i = x_{b_0} &= \frac{(n+4)^2}{n^2+12n+16} \end{cases} \tag{3.51}$$

Total opinion

$$x_{tot}^{(b_0)} = \frac{n\left(n^2+12n+24\right)}{2n^2+24n+32} \tag{3.52}$$

Now when all different possibilities of link placement have been found it is possible to compare them and determine which yields the the largest opinion gain and for what n this is the case. There is an assumption the the placement in a_0 is the optimal and the following calculations will prove that it is the case. Begin by compare a_0 and a. In this case, the total opinion, $\frac{n}{2}$, of the unmodified network is subtracted from the total opinion of the modified network to simplify the calculations.

$$x_{gain}^{(a)} = x_{tot}^{(a)} - \frac{n}{2} = \frac{n(n+6)(n-2)}{n^3+12n^2+24n-16} \tag{3.53}$$

$$x_{gain}^{(a_0)} = x_{tot}^{(a_0)} - \frac{n}{2} = \frac{3n^2+34n+48}{n^2+12n+16} \tag{3.54}$$

Then study the inequality

$$
\begin{aligned}
& x_{gain}^{(a_0)} > x_{gain}^{(a)} && (3.55)\\
\Leftrightarrow\ & \frac{3n^2+34n+48}{n^2+12n+16} > \frac{n(n+6)(n-2)}{n^3+12n^2+24n-16} && (3.56)\\
\Leftrightarrow\ & (3n^2+34n+48)(n^3+12n^2+24n-16) > n(n+6)(n-2)(n^2+12n+16) && (3.57)\\
\Leftrightarrow\ & 3n^5+70n^4+528n^3+1344n^2+608n-768 > n^5+16n^4+52n^3-80n^2-192n && (3.58)\\
\Leftrightarrow\ & 2n^5+54n^4+476n^3+1424n^2+800n-768 > 0 && (3.59)
\end{aligned}
$$

which is fulfilled for all $n \geq 1$. Next compare a_0 to b

$$
\begin{aligned}
x_{tot}^{(a_0)} > x_{tot}^{(b)} \Leftrightarrow\ & \frac{n^3+18n^2+84n+96}{2n^2+24n+32} > \frac{n(n+8)(n+4)}{2n^2+24n+56} && (3.60)\\
\Leftrightarrow\ & 12n^4+272n^3+2064n^2+5984n+5376 > 0 && (3.61)
\end{aligned}
$$

which holds for all $n \geq 0$. Finally compare a_0 and b_0

$$
\begin{aligned}
x_{tot}^{(a_0)} > x_{tot}^{(b_0)} \Leftrightarrow\ & \frac{n^3+18n^2+84n+96}{2n^2+24n+32} > \frac{n\left(n^2+12n+24\right)}{2n^2+24n+32} && (3.62)\\
\Leftrightarrow\ & 6n^2+60n+96 > 0 && (3.63)
\end{aligned}
$$

which is true for all $n \geq 0$. Thus it is proved that a_0 is the optimal placement in a barbell graph. □

4

A distributed algorithm for social influence maximization

In the previous chapter explicit solutions were introduced for a few specific network topographies. While this is useful for insight, it provides little benefit for broad application. A distributed algorithm working on any network would be ideal and this is the goal for this chapter. First, in Section 4.1, a formula providing a solution for general network types is presented and proved as well as an attempt for an intuitive understanding of it. Then, in Section 4.2, steps are made to perform a suitable distributed implementation as well as a few brief comments on computational complexity.

4.1 Derivation of general formula

As explained earlier, it would be valuable to find a general formula to find the optimal link placement for a maximization the total opinion distribution. The following theorem provides a formula which computes the total opinion gain when a directed link is added from a regular node i to a stubborn node in $\mathcal{S}_1$. The optimal placement can then be discovered if this is calculated for every node.

THEOREM 4.1
Consider the two networks $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ and $\tilde{\mathcal{G}}^{(i)} = (\mathcal{V}, \tilde{\mathcal{E}}^{(i)}, \tilde{W}^{(i)})$ so that $\tilde{\mathcal{E}}^{(i)} = (\mathcal{E} \cup (i, s_1))$. Where $s_1 \in \mathcal{S}_1 \neq \emptyset$, $\mathcal{S}_0 \neq \emptyset$ and $i \in \mathcal{R}$. If y^* and $\tilde{y}^{(i)}$ are the opinion equilibria in $\mathcal{G}$ and $\tilde{\mathcal{G}}^{(i)}$ respectively. Then the opinion gain between the two graphs is

$$\gamma_i = \mathbb{1}'\tilde{y}^{(i)} - \mathbb{1}'y^* = \frac{z_i^*(1 - y_i^*)}{w_i + K_{ii}} \tag{4.1}$$

where

$$y^* = Qy^* + b, \quad b = Bu, \quad K = (I - Q')^{-1}, \quad z = K\mathbb{1}, \quad w = W\mathbb{1} \tag{4.2}$$

Proof Let $K = (\mathrm{I} - Q')^{-1}$ and $b = Bu$ then

$$x = K'b \tag{4.3}$$
$$z = K\mathbb{1} \tag{4.4}$$

This assumes that $(I - Q')$ and $(I - Q)$ is invertible. Since Q fulfils the requirements of Lemma 2.1, then $(I - Q)$ is invertible. Then it follows that $(I - Q')$ is invertible from

$$((I - Q)^{-1})' = ((I - Q)')^{-1} = (I - Q')^{-1}. \tag{4.5}$$

If the sum of x is the total opinion in the network, then the relationship between opinion dynamics and network flow can be acquired as following

$$\mathbb{1}'y = \mathbb{1}'K'b = z'b. \tag{4.6}$$

For the network with an added directed link from node i the following changes are needed

$$\tilde{b}^{(i)} = b + \frac{1 - b_i}{w_i + 1}\delta^{(i)} \tag{4.7}$$
$$(\tilde{Q}^{(i)})' = Q' - \frac{Q'\delta^{(i)}(\delta^{(i)})'}{w_i + 1} \tag{4.8}$$

then K can be written and simplified, using the Sherman-Morrison formula presented in Appendix, as

$$\tilde{K}^{(i)} = \left(\mathrm{I} - Q' + \frac{Q'\delta^{(i)}(\delta^{(i)})'}{w_i+1}\right)^{-1} = K - \frac{KQ'\delta^{(i)}(\delta^{(i)})'K'}{\frac{w_i+1}{1+(\delta^{(i)})'K\frac{Q'\delta^{(i)}}{w_i+1}}} \tag{4.9}$$

$$= K - \frac{KQ'\delta^{(i)}(\delta^{(i)})'K'}{w_i+1+(\delta^{(i)})'KQ'\delta^{(i)}} = K - \frac{KQ'\delta^{(i)}(\delta^{(i)})'K'}{w_i+K_{ii}}. \tag{4.10}$$

Then finally

$$\tilde{z}^{(i)} = \tilde{K}^{(i)}\mathbb{1} = K\mathbb{1} - \frac{KQ'\delta^{(i)}(\delta^{(i)})'K'\mathbb{1}}{w_i+K_{ii}} = z - \frac{KQ'\delta^{(i)}z_i}{w_i+K_{ii}}. \tag{4.11}$$

The gain of a link added to node i is then computed through

$$\gamma_i = \mathbb{1}'\tilde{y}^{(i)} - \mathbb{1}'y = (\tilde{z}^{(i)})'\tilde{b}^{(i)} - z'b \tag{4.12}$$

$$= \frac{(1-b_i)z_i}{w_i+1} - \frac{z_i(Qy)_i}{w_i+K_i i} - \frac{z_i(1-b_i)(\delta(i))'KQ'\delta(i)}{(w_i+1)(w_i+K_{ii})} \tag{4.13}$$

$$= z_i\left[\frac{1-b_i}{w_i+1} - \frac{y_i-b_i}{w_i+K_{ii}} - \frac{(1-b_i)(K_{ii}-1)}{(w_i+1)(w_i+K_{ii})}\right] \tag{4.14}$$

$$= \frac{z_i(1-y_i)}{w_i+K_{ii}} \tag{4.15}$$

□

Looking at the resulting formula

$$\gamma_i = \frac{z_i(1-y_i)}{w_i+K_{ii}} \tag{4.16}$$

some thoughts can be given on the interpretation of the different components. Consider again a line graph with ten nodes, computation of the formula components for the regular nodes yields the results presented in Figure 4.1 An interesting observation is that all components are unbiased to the target opinion except $(1-y_1)$. It is intuitive that $(1-y_i)$ provides a larger gain if the

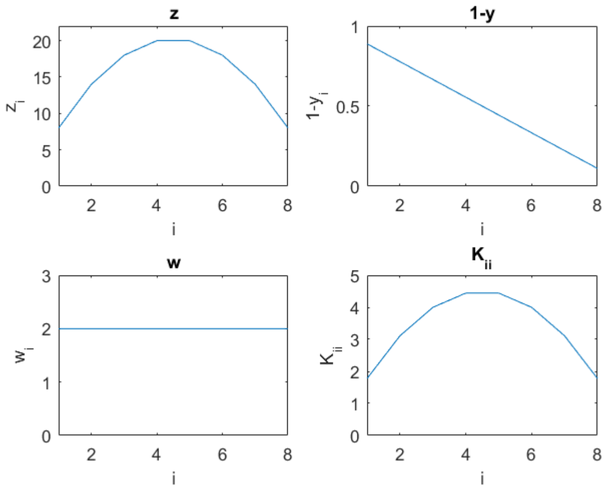


Figure 4.1 Computation of the algorithm components for regular nodes in a line graph with ten nodes.

previously held opinion y_i is smaller since there is more room for increase. The term z_i can be given an understanding by looking at its dynamics

$$z(t+1) = Q'z(t) + \mathbb{1} \tag{4.17}$$

with equilibrium

$$z^* = Q'z^* + \mathbb{1} \tag{4.18}$$

and compare it to the previously presented PageRank centrality (2.7). It is not directly translatable but it does give an indication that z_i is a measurement of centrality. The opinion gain is then normalized by w_i and K_{ii}. First, the out-degree w_i stems from the normalization of the added link is then intuitive to remain. From Theorem 2.4 it is known that K_{ii} can be translated to

$$K_{ii} = \frac{1}{1-p_i} \tag{4.19}$$

where p_i is the probability for a random walk starting in i to ever return to i instead leaving the network. Thus this term will be larger if the node i is placed further away from the stubborn nodes which also is given from the computation results. A larger gain will be provided if the node is closer to a stubborn node. Note that this K_{ii} and all other terms is calculated for the network before the directed link is added. All in all, everything except $(1-y_i)$ can be considered as a centrality measurement for the node i.

4.2 The quest for a distributed implementation

To use the algorithm in practice it is necessary to find z, y, w and K. From the graph definition, W, Q, B and u are already provided. Then make the following definitions

$$w = W\mathbb{1} \tag{4.20}$$

$$b = Bu. \tag{4.21}$$

where w is earlier known as the vector where every element w_i is the out-degree for node i. The term y^* is the equilibrium of the previously used opinion dynamics

$$y(t+1) = Qy(t) + b \tag{4.22}$$

and can then be calculated by solving the linear system

$$(\mathrm{I} - Q)y^* = b \tag{4.23}$$

usually done by the use of a Gaussian elimination based solver. The computational complexity of this is of order $\mathcal{O}(n^3)$. In general, sparse matrices are used which reduces the complexity. However this is still not optimal for larger matrices and it is not distributed. Instead an iterative approximation can be found by simply iterating the opinion dynamics, reducing the computational complexity to $\mathcal{O}(|\mathcal{E}|)$ for each step. While this is likely to be slower because of a large number of needed iterations, it is less prone to numerical errors and above all it is distributed. The approximation of z^* can be done in a similar way. While it also can be computed using a Gaussian elimination based solver, the same benefits of a distributed iterative method applies here. The dynamics for z is then

$$z(t+1) = Q'z(t) + \mathbb{1} \tag{4.24}$$

for some $z(0) \in \mathbb{R}_+^{\mathcal{R}}$. As the opinion dynamics, this can be assumed to have converged sufficiently close to z^* after a limited amount of iterations.

4.3 Distributed computation of K_{ii}

To calculate the K-matrix exactly it requires the computation of the inverse of $(I-Q')$. This generally results in numerical errors, especially if $(I-Q')$ is close to singular. Since only the diagonal of K is needed, this can be given another interpretation. As shown in Theorem 2.4, K_{ii} can be interpreted as

$$K_{ii} = ((I-Q')^{-1})_{ii} = \sum_{k=0}^{\infty}((Q')^k)_{ii} = 1 + \frac{p_i}{1-p_i} = \frac{1}{1-p_i} \tag{4.25}$$

where p_i is the probability for a random walk, starting in i, will ever return to the origin i before leaving the network. This probability is found by simulating random walks on the network, by starting in node i and count every time it has returned to itself and every time it has exited the network by hitting a stubborn node. Then the probability will be

$$p_i = \frac{N_{\text{returns}}}{N_{\text{returns}} + N_{\text{exits}}} \tag{4.26}$$

where N_{returns} is the number of returns and N_{exits} the number of exits. As a random walk can take a lot of time to return or exit the network it can for example be considered to have left the network after a certain amount of steps.

Alternatives for K-approximation

A common approach to estimate the matrix inverse is first to rewrite it using the power series expansion

$$K = \left(\mathrm{I} - Q'\right)^{-1} = \sum_{k=0}^{\infty}(Q')^k \tag{4.27}$$

which indicates that a sufficient approximation of K can be made by truncating this power series after a sufficiently large amount of iterations. This results in the estimation

$$K_{est} = \sum_{k=0}^{N} (Q')^k \tag{4.28}$$

where N is the number of iterations. The direct implementation is very computationally heavy as the computational cost is $\mathcal{O}(N(N+1)n^3)$. However, a simplification can be made by using the exponential $(Q')^k$ when computing $(Q')^{k+1}$. This results in the following iterative scheme

$$(Q')^k = (Q')^{k-1}(Q'+I), \quad 1 \leq k \leq N. \tag{4.29}$$

This reduces the computational cost to $\mathcal{O}(Nn^3)$. Both of these methods are non-distributed and should not be considered.

5

Numerical simulations

In this chapter, simulations are done using the algorithm presented in the previous chapter. First, in Section 5.1, a description of the different network topologies are presented as well as the placement of stubborn nodes. To verify the results of simulations on these networks, an approximation of the real solution is needed. These methods are presented in Section 5.2. Finally, in Section 5.3 simulations are done and the results are presented along with a comparison of the approximated real solution.

5.1 Data sets

The algorithm was implemented on a few networks. First a few artificial networks of varying sizes were chosen and implemented, these include the now familiar line graph and a toroid. A toroid with $n = 144$ nodes can be seen in Figure 5.1. The two stubborn nodes are placed in the same circle around the toroid in such a way that if there are $n_1 = 12$ nodes in that that circle, the stubborn nodes have position 1 and $\frac{n_1}{2} = 6$. In this kind of toroid, one line circling the center of the toroid have the same number of nodes that circles a line around the side. This result in the total number of nodes being a square number. The toroid used in the simulations have a total of 100 nodes numbered around the center, 1-10 in the first line, 11-20 in the second etcetera.

A real dataset were acquired from SNAP [Leskovec and Krevl, 2014] and include a part from Facebook with $n = 4039$. As stubborn nodes were not present, they had to be added and 30 of the nodes with the highest indegree

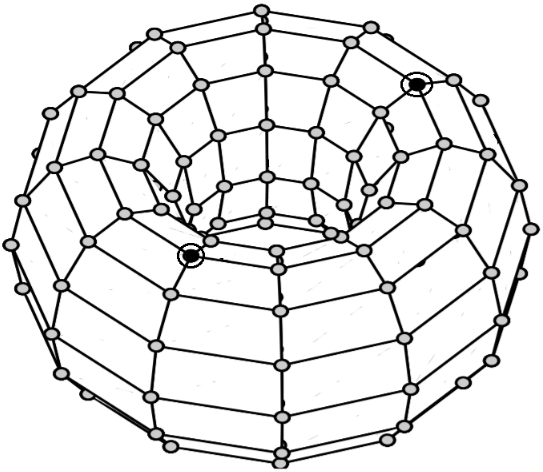


Figure 5.1 Toroid with $n = 144$ nodes and two stubborn nodes.[Acemoğlu et al., 2013]

were converted into stubborn nodes. In total the algorithm was used on the following networks:

- Line graph with 10 nodes
- Line graph with 50 nodes
- A toroid with 100 nodes
- Facebook network with 4039 nodes.

5.2 Verification of results

To be able to verify results from running the algorithm on actual networks, it is important to have at least an approximate answer to compare with. For the line graph this solution is explicitly known from Chapter 3. For the toroid, the matrix Q is assumed to be far enough from singular to be able to solve the linear system

$$(I - \tilde{Q}^{(i)})\tilde{x}^{(i)*} = \tilde{B}^{(i)}u \tag{5.1}$$

with a sufficiently small enough numerical error. This is done for every i to get the opinion equilibrium for every possible positioning of the tail of the added link. Then

$$\sum_j \tilde{x}_j^{(i)*} - \sum_j x_j^* \tag{5.2}$$

yields the opinion gain when a the tail of the directed link is placed in i. As earlier, x^*, is the opinion equilibrium for the unmodified network. For the Facebook network, the opinion dynamics

$$x(t+1) = Qx(t) + Bu, \quad 0 \leq t \leq 1000 \tag{5.3}$$

is used for all choices of the modified Q and B matrix.

5.3 Simulations

For the networks, the dependence on two different parameters in the algorithm were studied. They are presented as *steps* and *K-steps*. Steps are the number of iterations that are done in the approximation of x and z described in (4.22) and (4.24). The method used to estimate K_{ii} is the random walk method described in Section 4.3. The number of iterations, K-steps, in this case is the least amount of random walks performed for each node. To calculate the relative error the following formula is used

$$e_r = \frac{||\gamma_{sim} - \gamma_{real}||}{||\gamma_{real}||} \tag{5.4}$$

where γ_{sim} is the simulated gain for every node and γ_{real} the verification value described in the former section

Line graph with 10 nodes

Figure 5.3 depicts how the simulated gain varies with different choices of K-steps when the number of normal steps is set to 10000. In Figure 5.2 the

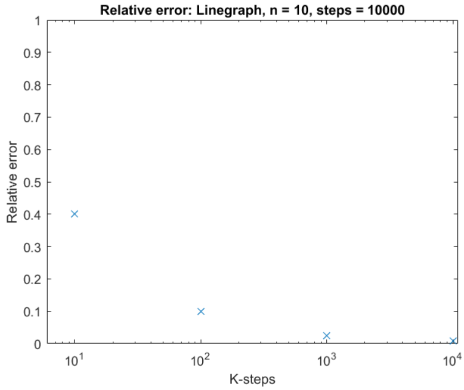


Figure 5.2 Relative error in a line graph with 10 nodes depending on regular node. Number of K-steps varying from 10 to 10000. Amount of steps is set to 10000.

relative error versus number of K-steps is presented. Even though ten K-steps provides a large error, the algorithm shows a maximum in node 2, which is the optimal placement. However this result is unlikely to be consistent considering the stochastic nature of the random walk method. Then at least 1000 K-steps seems to provide a reliable solution.

Conversely, in Figure 5.4, the number of K-steps is held constant at 10000 and the amount of steps is the variable to be studied. Figure 5.5 shows the relative error in this case. In this case 100 steps and above seems to provide an accurate result.

Line graph with 50 nodes

The same is done for a line graph with 50 nodes. First, in Figure 5.6, the gain for every node when the number of K-steps is varied and steps held constant at 10000. The relative error is presented in Figure 5.7. When the number of nodes is increased, the unreliability of the random walk method for few iterations is more apparent. At least 10000 K-steps should be used, perhaps even more since the maximum of the simulation does not coincide with the maximum of the true gain. Then, K-steps is held constant at 10000 and the total gain for every node is presented in Figure 5.8 for different number of steps.

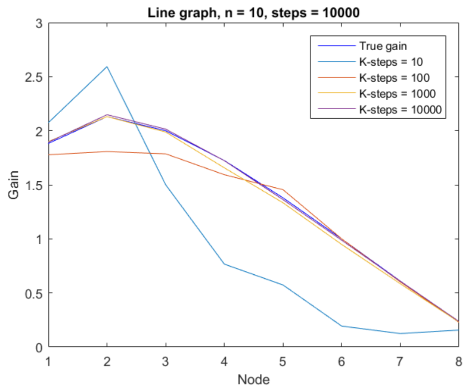


Figure 5.3 Gain in a line graph with 10 nodes depending on regular node. Number of K-steps varying from 10 to 10000. Amount of steps is set to 10000.

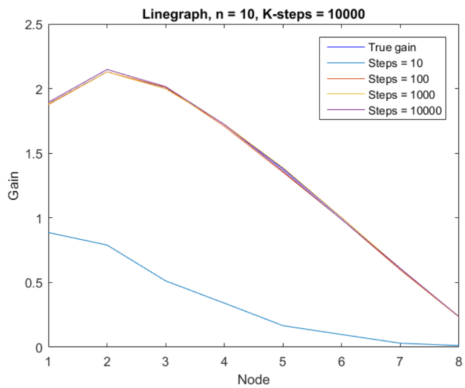


Figure 5.4 Gain in a line graph with 10 nodes depending on regular node. Number of steps varying from 10 to 10000. Amount of K-steps is set to 10000.

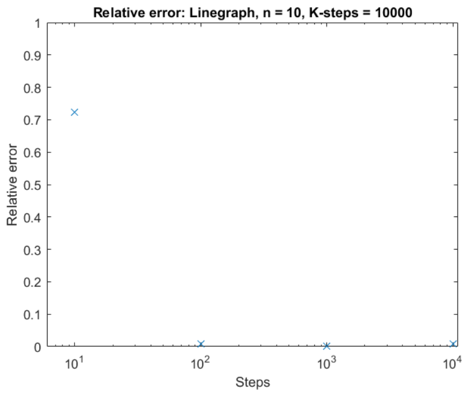


Figure 5.5 Relative error in a line graph with 10 nodes depending on regular node. Number of steps varying from 10 to 10000. Amount of K-steps is set to 10000.

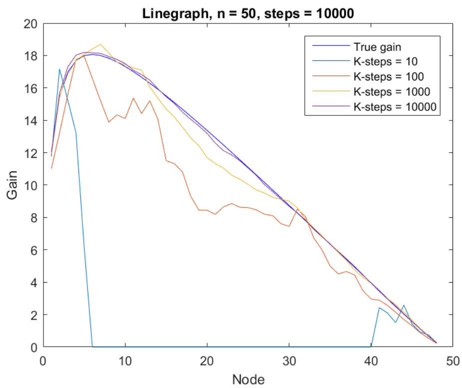


Figure 5.6 Gain in a line graph with 50 nodes depending on regular node. Number of K-steps varying from 10 to 10000. Amount of steps is set to 10000.

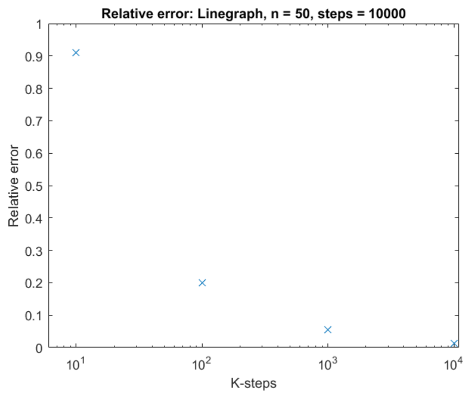


Figure 5.7 Relative error in a line graph with 50 nodes depending on regular node. Number of K-steps varying from 10 to 10000. Amount of steps is set to 10000.

The relative error can be seen in Figure 5.9. In this case, at least 10000 steps is needed for the algorithm to be close to the true gain. Again, the maximum does not coincide perfectly with the true gain. If this remains consistent this could be an indication of an implementation error.

Toroid

Same thing is done for a toroid with varying K-steps in Figure 5.10 and corresponding relative error in Figure 5.11 as well as with varying number of steps in Figure 5.12 and relative error in Figure 5.13. As the node numbering is not as intuitive as the line graph, the figures depicting gain as a function of node numbers can be considered less important. These results gives a few interesting observations, even though the number of nodes is twice that of the line graph in the previous section, the convergence seems to be faster. At just 1000 steps and the equal amount of K-steps, the error is considerably small. Thus, the number of needed iterations does not only depend on network size, but also how close to each other the nodes are. In the line graph, the longest path between two nodes is around 50, while in the toroid it is around 20.

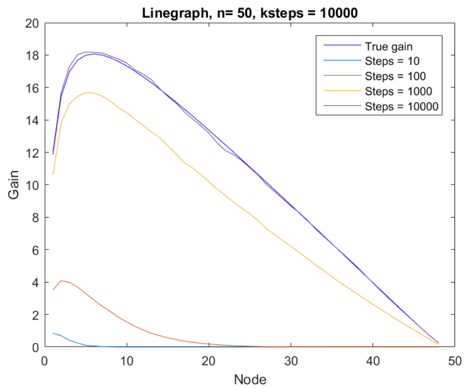


Figure 5.8 Gain in a line graph with 50 nodes depending on regular node. Number of steps varying from 10 to 10000. Amount of K-steps is set to 10000.

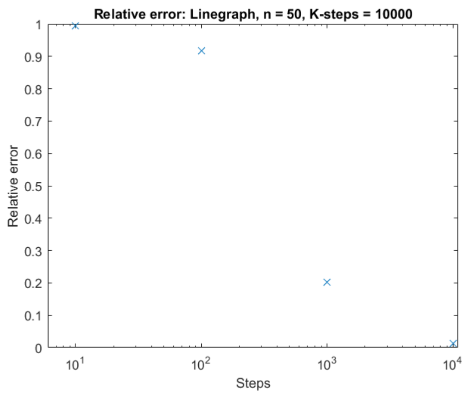


Figure 5.9 Relative error in a line graph with 10 nodes depending on regular node. Number of steps varying from 10 to 10000. Amount of K-steps is set to 10000.

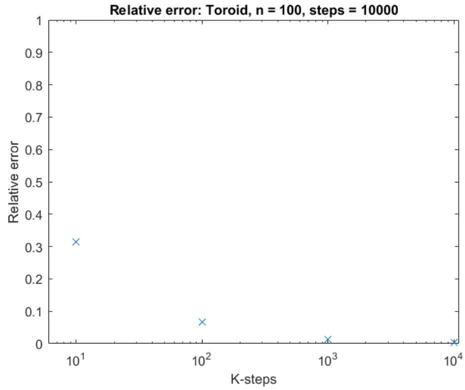


Figure 5.11 Relative error as a function of number of K-steps in a toroid with 100 nodes. Normal steps is set constantly to 10000.

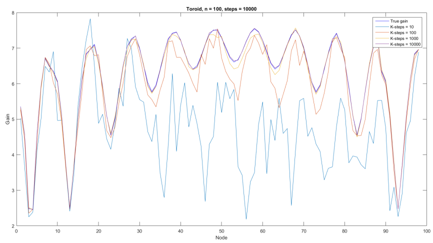


Figure 5.10 Gain for every regular node in a toroid with 100 nodes. Amount of K-steps is set to 10000.

Facebook

Finally, simulations were done for the Facebook-network. This is more important as it is more of a real world example. Then the previous simulations can be considered as a verification that the algorithm is working. Figure 5.14 provides a comparison between the simulated value along with the verification value described earlier. Observing this it is apparent that most options

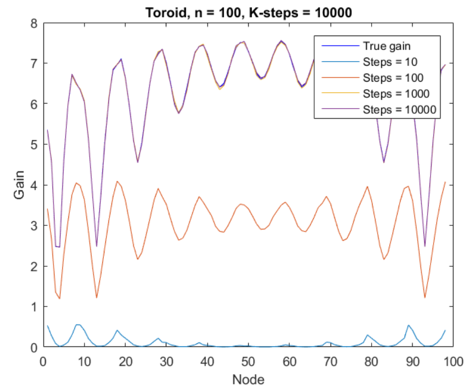


Figure 5.12 Gain for every regular node in a toroid with 100 nodes. Amount of K-steps is set to 10000.

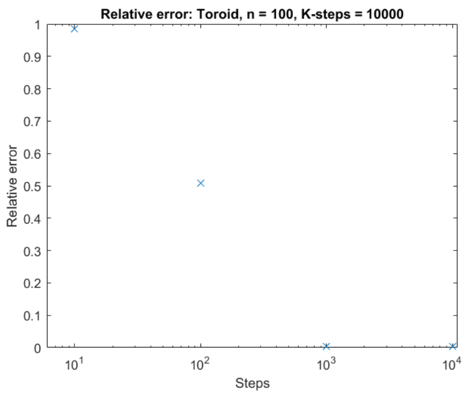


Figure 5.13 Relative error as a function of number of steps in a toroid with 100 nodes. K-steps is set constantly to 10000.

do not change the total opinion considerably but there are some nodes that do, for example the nodes numbered above 4000. The optimal solution in this case would provide an opinion gain of 22, which is equal to an total opinion increase by 1.5%. The relative error as a function of step numbers is provided in Figure 5.15. The reason why it is enough with just ten K-steps is because depending on the positioning in the network of nodes, some will take a longer time to reach a total number of ten exits and returns. In this time, many other nodes have reach several thousand, why it still provides high enough accuracy.

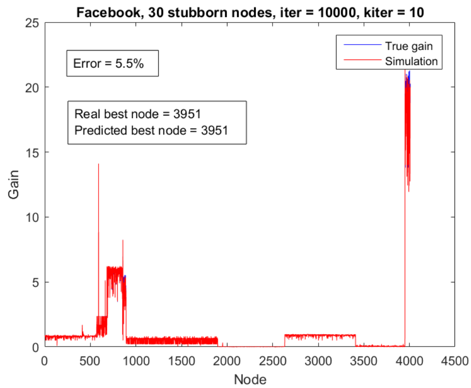


Figure 5.14 Simulated and real gain in the Facebook-network.

Looking at the computation times provided in Figure 5.16, the results are confusing. While it is apparent that the K_{ii} approximation contributes most to the computation time. When increasing the number of steps for x and z approximation, the total computation time actually decreases, which does not make sense. As this remains consistent for several simulation runs, I want to attribute this to coding error. Other unlikely reasons could be that Matlab use previously done computations to refrain from repetitions. To compare with the truncated power series method, simulations were done for varying truncations lengths. The relative error is provided in Figure 5.17 and it is quickly clear that 1000 iterations is not close to enough and the error is almost 100%. The computation time is also considerably large, see Figure 5.18, which does not make it a viable option.

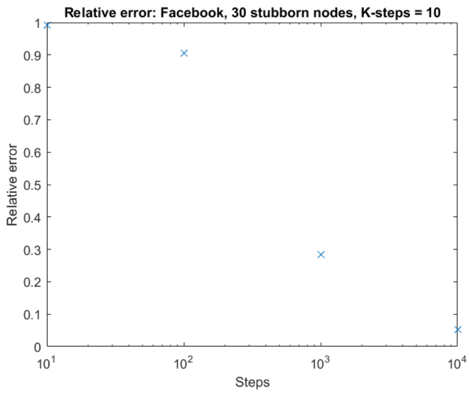


Figure 5.15 Relative error for the Facebook-network as a function of number of steps. K-steps is set to 10.

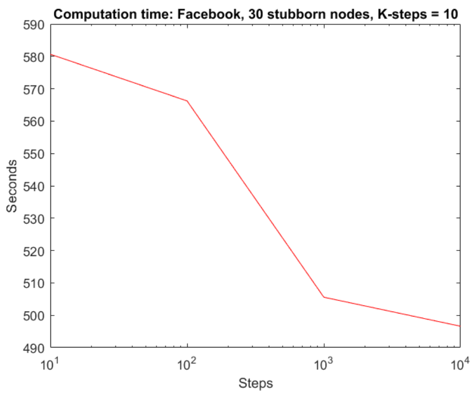


Figure 5.16 Computation time for the whole algorithm when varying the number of steps.

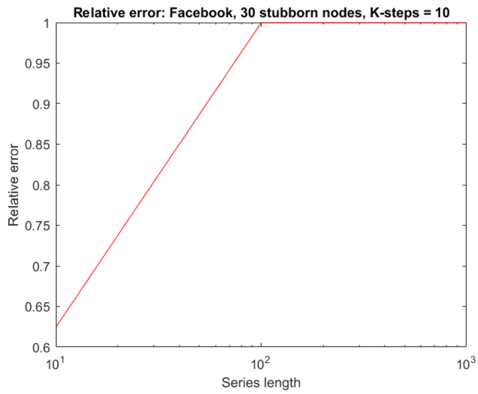


Figure 5.17 Relative error when using the truncated power series method.

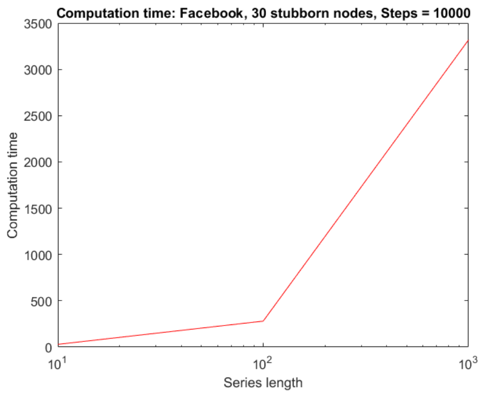


Figure 5.18 Computation time for the whole algorithm when using truncated power series method. The x-axis shows series length.

6

Conclusion

The objective in this thesis was to find the optimal placement of a directed link in order to maximize the resulting influence in the network. It has been shown that the complexity of this problem can be reduced by always letting the directed link point to a stubborn node of desired opinion. The result is that it is only necessary to find the node from which the directed link should stem. We then progressed into finding the optimal tail placement for the line graph and the barbell graph as well as the case when multiple lines are present.

Next, we formulated a formula, for general network topographies, to compute the induced gain brought by the addition of a directed link. We proceed by introducing a distributed algorithm using this formula. Simulations shows the speed of convergence in a few different network types.

The benefit that a distributed implementation has been found working on general graphs cannot be stressed enough. A distributed implementation increases the possibilities for actual use. If again we compare to the similar problem of turning a node into a stubborn node, this was not fully achieved in the works studied.

A real life implementation of the algorithm raises further questions. If we want to add many links, what is the optimal approach? Should we run the algorithm once and use the best nodes or should one instead use an iterative approach by adding one link and then run the algorithm again? Perhaps it is possible to find a middle ground in this case. Other issues is how we can quantify the opinion of people. And how can we quantify the strength of social connections? Future work could try and answer these questions.

There are also some things not modelled in this thesis that should be taken in consideration, such as the degree of self confidence of a node. This can be modelled by adding a self loop to every node so that their own opinion is taken into consideration in the update process.

Bibliography

Acemoglu, D. and A. Ozdaglar (2011). "Opinion dynamics and learning in social networks". *Dynamic Games and Applications* **1**:1, pp. 3–49.

Acemoğlu, D., G. Como, F. Fagnani, and A. Ozdaglar (2013). "Opinion fluctuations and disagreement in social networks". *Mathematics of Operations Research* **38**:1, pp. 1–27. DOI: `10.1287/moor.1120.0570`.

Bonacich, P. (1987). "Power and centrality: a family of measures". *American Journal of Sociology* **92**:5, pp. 1170–1182.

Brin, S. and L. Page (1998). "The anatomy of a large-scale hypertextual web search engine". In: *Seventh International World-Wide Web Conference (WWW 1998)*. URL: `http://ilpubs.stanford.edu:8090/361/`.

Como, G. and F. Fagnani (2010). "Scaling limits for continuous opinion dynamics systems". *ArXiv e-prints*. arXiv: `1003.3384 [math.PR]`.

Como, G. and F. Fagnani (2016). "From local averaging to emergent global behaviors: the role of network interconnections". *Systems & Control Letters* **95**, p. 71.

Degroot, M. H. (1974). "Reaching a consensus". *Journal of the American Statistical Association* **69**:345, pp. 118–121.

Friedkin, N. E. (2015). "The problem of social control and coordination of complex systems in sociology: a look at the community cleavage problem". *IEEE Control Systems* **35**:3, pp. 40–51.

Friedkin, N. (2010). "A formal theory of social power". **12**, pp. 103–126.

Gravner, J. (2014). *Lecture Notes for Introductory Probability*. Mathematics Department, University of California, pp. 153–155.

Harary, F. (1959). "A criterion for unanimity in french's theory of social power", pp. 168–182.

Holst, A. and V. Ufnarovski (2014). *Matrix Theory*. Studentlitteratur AB, pp. 267–270.

Kempe, D., J. Kleinberg, and E. Tardos (2003). "Maximizing the spread of influence through a social network". In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '03. Washington, D.C., pp. 137–146. ISBN: 1-58113-737-0.

Leskovec, J. and A. Krevl (2014). *SNAP Datasets: Stanford large network dataset collection*. `http://snap.stanford.edu/data`.

Rossi, W. S. and P. Frasca (2016). "The harmonic influence in social networks and its distributed computation by message passing". *ArXiv:1611.02955v1*.

Sherman, J. and W. J. Morrison (1950). "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix". *Ann. Math. Statist.* **21**:1, pp. 124–127. DOI: `10.1214/aoms/1177729893`.

Vassio, L., F. Fagnani, P. Frasca, and A. Ozdaglar (2014). "Message passing optimization of harmonic influence centrality". *arXiv:1310.0005v2 [math.OC]*.

Appendix

The Sherman-Morrison formula

The Sherman-Morrison formula [Sherman and Morrison, 1950] gives an expression for how the inverse changes of a matrix when one element in said matrix has been changed. Consider the invertible matrix A with inverse A^{-1}. Let

$$B = A + \Delta \tag{6.1}$$

$$\Delta = a\delta^{(i)}\delta^{(i)\prime}, \quad a \in \mathbb{R} \tag{6.2}$$

where a can be interpreted as the change of index (i, j) in the matrix A. Then the inverse of B can be written as

$$B^{-1} = (A + \Delta)^{-1} = A^{-1} - \frac{aA^{-1}\delta^{(i)}\delta^{(i)\prime}A^{-1}}{1 + a\delta^{(i)\prime}A^{-1}\delta^{(i)}} \tag{6.3}$$

Lund University
Department of Automatic Control
Box 118
SE-221 00 Lund Sweden

Document name
MASTER'S THESIS

Date of issue
October 2017

Document Number
TFRT-6045

Author(s)
Joakim Guth

Supervisor
Giacomo Como, Dept. of Automatic Control, Lund University, Sweden
Gustav Nilsson, Dept. of Automatic Control, Lund University, Sweden
Anders Rantzer, Dept. of Automatic Control, Lund University, Sweden (examiner)

Title and subtitle
On Distributed Maximization of Influence in Social Networks

Abstract

This thesis studies the problem of finding the optimal placement of a directed link in a graph representation of a social network in order to maximize the induced gain of the opinion equilibrium. The model assumes the presence of a set of stubborn nodes and applies a standard DeGroot opinion dynamics model. First we show that an added directed link should point to a stubborn node in order to maximize the impact of the link. The resulting problem reduction then allows for explicit solutions of where the directed link should origin in a few common network topographies such as the line graph and the barbell graph. A formula for the optimal tail placement for general graphs is then presented along with a distributed algorithm. Implementation and simulation are then performed again first on a few common network types and then on a small sub-network of Facebook.

Keywords

Classification system and/or index terms (if any)

Supplementary bibliographical information

ISSN and key title
0280-5316

ISBN

Language
English

Number of pages
1-64

Recipient's notes

Security classification

http://www.control.lth.se/publications/