Control and Prediction of Beliefs on Social Network

Tian Wang
Department of Physics
North Carolina State University
Email: twang9@ncsu.edu

Hamid Krim
Department of Electrical and Computer Engineering
North Carolina State University
Email: ahk@ncsu.edu

***Abstract*—In this paper we propose a belief flow model for social networks and evaluate its application on estimation of public converged beliefs. The model reveals that the control of beliefs in a social network heavily depends on its degree centralities and clustering coefficients. The application of this model to social network belief flow simulation leads to a capacity to control and predict the converged beliefs in a social network. Two different network models, preferential attachment model and generalized Markov Graph model, are applied to the belief flow model. Experiments with published real social network data are provided and demonstrate very good performance of the belief flow model as well as a comparison between different network models.**

***Index Terms*—Social Networks, Information Flow, Machine Learning**

I. Introduction to the model

The beliefs in a social network may have value for members of the network or even outsiders [4] [7]. For example, in a clinical study, a doctor may be interested in researching ways to influence patients' behaviour by "facilitating" interactions among patients. Consequently, prediction [1] [6] [3] or even control of the beliefs in a social network can be an important and interesting problem. As such, it requires a mathematical model to simulate, analyse the flow of beliefs, as well as optimize the control strategy, in a social network.

A. Basic Concepts and Definitions

A linear belief flow model includes a network, people's private beliefs in the network, updated beliefs at each time step, and a control strategy to shape beliefs. The model takes network, private beliefs and control strategy as inputs and final converged beliefs as outputs. The definitions of these concepts are introduced in the following paragraphs.

1) Network: The structure of a network plays a very important role as it influences the result of the final converged beliefs as well as the strategy of belief control. For a network G with N nodes, we use indices $i \in \{1, 2, ..., N\}$ to represent the nodes, and the set of nodes is defined as: $node_G = \{1, 2, ..., N\}$. The set of edges, $edge_G$, includes all pairs of connected nodes, $\{i, j\} \in edge_G$, in the network and the network is thus defined as: $G = \{node_G, edge_G\}$. The information of the network can also be represented by its adjacency matrix: $\{A\}$, whose elements are A_{ij}. And in this paper, we only focus on undirected binary networks.

In practice, the information of the network may not be complete, which means $\{A\}$ is not always available. Furthermore, the network may contain a large number of nodes or edges, which requires an expensive computational power to process. To solve such problems, network models are necessary. A good network model can help calculate the converged beliefs using less information than $\{A\}$, and more efficiently. Two important network models, the preferential attachment model [2] and the generalized Markov Graph model [9], will be introduced and applied in the analysis of belief flow .

2) Belief: We invoke two kinds of beliefs in this model: private belief and updated belief. The former is unchanged and taken as an input of the model. The latter, however, updates at each time step and will be converging to a limit.

a) Private beliefs: Private beliefs abstract the intrinsic characteristics of nodes in a network. They will not be changed during the process of information flow. In this model, node i in the network takes the private belief as a random number $w_i \in [-1, 1]$ with distribution $p(w_i)$. The distribution $p(w_i)$ is common knowledge to everyone in the network.

b) Current beliefs: A current Belief $B_{i,T}$ describes the current opinion of node i in a network at time step T. It lies in the range $[-1, 1]$. For an arbitrary node i in the network, $B_{i,T}$ can be observed by its neighbour and is initialized as a private belief w_i, i.e., $B_{i,0} = w_i$. The value of $B_{i,T}$ is determined as the average of the current beliefs of the neighbours of node i and the private belief of node i, w_i. $B_{i,T}$ will be updated at each time step and will converge to a limit $B_{i,\infty}$ in certain networks, as will be explained in Section II

3) Control Power: Control power is used to show how much the beliefs in a network are altered from their initial status. Control power over an arbitrary node i is defined as the difference between the converged belief $B_{i,\infty}$ and the initial belief w_i: $c_{p_i} = B_{i,\infty} - w_i$. And the averaged c_{p_i} all over the network is thus the *network control power*: $c_p = \Sigma_{i=1}^{N} c_{p_i}/N$.

4) Control Strategy: To control the overall behaviour of the network, we propose a control strategy which chooses certain people in the network, so called control nodes, and asks them to broadcast chosen beliefs to their neighbours. The set of control nodes is defined as $\mathbf{C}$, with a cardinality c. The c control nodes are set, without loss of generality, as the first c nodes in the node set $node_G$. And the belief chosen to be broadcast by the i^{th} control nodes is the controlled belief C_i,

978-1-4673-3146-3/13/$31.00 ©2013IEEE

where $C_i \in [-1, 1]$. The choice of the control nodes depends on the structure of the network to reach the maximum control power.

B. Belief Flow Model

1) Current Belief Updated without Control Strategy: If the control strategy is not applied, at an arbitrary time step T, where $T \in \mathbb{Z}$, the current belief $B_{i,T}$ of node i is updated according to the average of the current beliefs of its neighbours at the previous time step, $\{B_{j,T-1} | \{i, j\} \in edge_G\}$, and its private belief w_i, as shown in Eq.(1):

$$B_{i,T} = \frac{\Sigma_{j,\{i,j\} \in edge_G} B_{j,T-1} + w_i}{d_i + 1}, \qquad (1)$$

where d_i is the degree of node i , and $B_{i,0}$ is initialized to 0.

2) Current Belief Updated with Control Strategy: To better show how the converged beliefs across the network G are influenced by the control strategy, we define the following concepts. The current belief vector is defined as $\mathbf{B}(T)$, whose elements are $B_{i,T}$, to represent the beliefs of the nodes. The time step $T \in \mathbb{N}$, and $B_{i,0}$ is set to w_i. The adjusted private belief vector is defined as: $\mathbf{w}^*$ whose elements are $w_i/(d_i + 1)$. And the adjusted adjacency matrix A^* contains elements $A^*_{i,j} = A_{i,j}/(1 + d_j)$. The control matrix is defined as M, where $M_{i,i} = 1$ if $i \notin \mathbf{C}$ and $M_{i,j} = 0$ otherwise. The control vector is V, where $V_i = C_i$ if $i \in \mathbf{C}$ and $V_i = 0$ otherwise. The updating process of the current belief vector $\mathbf{B}(T)$ is shown in Fig.(1),

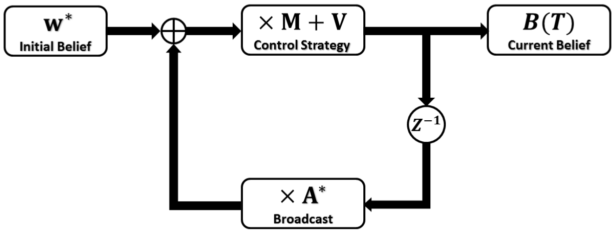

Fig. 1: Update of current belief.

The adjusted initial belief $\mathbf{w}^*$ and adjusted adjacency matrix A^* are for the calculation of averaged belief for each node. The control matrix M and control vector V are used to set the belief of control nodes as their corresponding controlled belief C_i. And Z^{-1} means the belief are updated based on the information of last time step. Eq.(2) shows the formula to calculate $\mathbf{B}(T)$:

$$\mathbf{B}(T) = [\mathbf{w}^* \times M + V] \times [\Sigma_{t=0}^{T-1} (A^* \times M)^t]. \qquad (2)$$

And if the summation $\Sigma_{t=0}^{T} (A^* \times M)$ in Eq.(2) converges to a finite matrix when T approaches ∞, the converged belief $\mathbf{B}(\infty)$ can be represented as in Eq.(3):

$$\mathbf{B}(\infty) = [\mathbf{w}^* \times M + V] \times [I - A^* \times M]^{-1}. \qquad (3)$$

II. Converged Beliefs Estimation

According to Eq.(3), to calculate the exact solution of converged belief vector $\mathbf{B}(\infty)$, the complete information of A is needed. In addition, the computational cost is the inversion of matrix $I - A^* \times M$. Such an exact solution doesn't particularly shed any light on the choice of control set $\mathbf{C}$ or the convergence speed of $\mathbf{B}(T)$ towards $\mathbf{B}(\infty)$. In order to reduce the information needed to predict the converged belief vector $\mathbf{B}(\infty)$, as well as provide detailed analysis about the control strategy and convergence speed, social network models are needed.

Two network models, preferential attachment model [2] and generalized Markov graph model [9], will be applied. The reason to choose these two models is that they both provide probabilistic properties about the element $A_{i,j}$ of adjacency matrix A. The preferential attachment model assumes that $A_{i,j}$ only depend on the degree of nodes i and j. The generalized Markov graph model, on the other hand, extends the dependence of $A_{i,j}$ to both degree and clustering coefficient of nodes i and j.

A. Preferential Attachment Model

1) Basic assumption: One of the basic assumptions of the preferential attachment model is that the probability of a node i attached by a new edge is proportional to its degree d_i [2]:

$$\frac{\partial d_i}{\partial t} \sim d_i. \qquad (4)$$

Based on this assumption, we can derive the probability $P_{i,j}$ of an edge established between nodes i and j, as shown in Theorem II.1 [10]. ${}^1P_{i,j}$ will play an important role in the analysis of control power estimation, and of control strategy as it represents the information of adjacency matrix A.

Theorem II.1. *The probability ${}^1P_{i,j}$ of two nodes i and j connected in network G is:*

$${}^1P_{i,j} = \frac{d_i d_j}{\Sigma_{k=1}^{N} d_k}, \qquad (5)$$

where d_k is the degree corresponding to node k.

2) Calculation of Control Power: Theorem II.1 reveals the statistical properties of adjacency matrix A. If we take A as a random matrix, combined with Eq.(2) and Eq.(3), we are able to give the expected value of converged belief for all the nodes in the network, which is shown in Theorem II.2 [10].

Theorem II.2. *In a preferential attachment model, the expected value of converged belief ${}^1B_{i,\infty}$ of a non-controlled node i, $i \notin \mathbf{C}$:*

$$\overline{{}^1B}_{i,\infty} = \frac{1}{\Sigma_{k=1}^{N} d_k} \frac{d_i}{1 + d_i} \frac{\Sigma_{j=1}^{c} C_j d_j + \Sigma_{j=c+1}^{N} \frac{w_j}{1+d_j} d_j}{1 - \beta_1}, \qquad (6)$$

where w_j is the private belief of node i, m is the average number of edges in a network G with N nodes, d_i is the degree corresponding to node i, C_j is the controlled belief of control node j, c is the number of control nodes, and β_1 is a constant which is smaller than 1: $\beta_1 = \Sigma_{k=c+1}^{N} \frac{{d_k}^2}{1+d_k} \diagdown \Sigma_{k=1}^{N} d_k$.

And the expected value of control power $\overline{{}^1c_{p_i}}$ will be: $\overline{{}^1B_{i,\infty}} - w_i$. According to Eq.(6), we can see that the control strategy has a direct impact on the converged belief. If controlled beliefs C_i are fixed, then the maximization of $|{}^1c_{p_i}|$ requires the selection of a control group $\mathbf{C}$ to include nodes with highest degrees in the network G [10].

The information needed for the calculation in Theorem II.2 is the degree list of network G, which is far less than the information of adjacency matrix A. And the computational cost of such calculation is $O(N)$, which is much more efficient than the matrix inverse calculation required by Eq.(3).

B. Generalized Markov Graph Model

1) Basic assumption: In Wang and Krim [9], the generalized Markov Graph model is introduced as a natural extension of preferential attachment model. In such a model, the probabilistic dependence on an edge is extended from the other attached edges to attached triangles. As degree is used to describe the dependence on attached edges, the clustering coefficient [9], which is related to both edges and triangles, is added to the description of dependence in a generalized Markov Graph model. The assumption about the probability of a node i attached by a new edge in a generalized Markov Graph model then becomes:

$$\frac{\partial d_i}{\partial t} \sim d_i(1+\gamma_i)^\alpha, \tag{7}$$

where d_i is degree of node i, γ_i is the clustering coefficient of node i, and α, which is called clustering weight, is determined by the property of the network G. To prevent zero clustering coefficient from making probability of a node getting an edge vanish, we use $(1+\gamma_i)$ instead of γ_i.

Based on Theorem II.1, we add the dependence of clustering coefficient. And to make the summation of ${}^2P_{i,j}$ still be the sum of degrees $\Sigma_{k=1}^N d_k$, the probability ${}^2P_{i,j}$ becomes [10] :

$${}^2P_{i,j} = \frac{d_i(1+\gamma_i)^\alpha d_j(1+\gamma_j)^\alpha}{\eta} \Sigma_{k=1}^N d_k, \tag{8}$$

where η is:

$$\eta = \Sigma_{i=1}^N \Sigma_{j=1, j\neq i}^N d_i(1+\gamma_i)^\alpha d_j(1+\gamma_j)^\alpha.$$

2) Calculation of Control Power: Based on Eq.(8), we developed the converged belief for the generalized Markov Graph model, as shown in Theorem II.3 [10].

Theorem II.3. *Define constant β_2 as*

$$\beta_2 = \frac{\Sigma_{k=c+1}^N \frac{(d_k(1+\gamma_k)^\alpha)^2}{1+d_k}}{\eta} \Sigma_{k=1}^N d_k. \tag{9}$$

If $|\beta_2| < 1$, in a generalized Markov Graph model, the expected value of converged belief ${}^2B_{i,\infty}$ of a non-controlled nodes i, $i \notin \mathbf{C}$ is:

$$\overline{{}^2B_{i,\infty}} = \frac{\Sigma_{k=1}^N d_k}{\eta} \frac{d_i(1+\gamma_i)^\alpha}{1+d_i} \frac{\Sigma_{j=1}^c C_j d_j(1+\gamma_j)^\alpha + \Sigma_{j=c+1}^N \frac{w_j}{1+d_j} d_j(1+\gamma_j)^\alpha}{1-\beta_2}, \tag{10}$$

where w_j is the private belief of node i in a network G with N nodes, d_i is the degree corresponding to node i, γ_i is the clustering coefficient of node i, C_j is the controlled belief of control node j, c is the number of control nodes, α is the clustering weight for network G, η is defined in Eq.(8).

The expected control power $\overline{{}^2c_{p_i}}$ will be: $\overline{{}^2B_{i,\infty}} - w_i$. According to Eq.(9), the clustering coefficient list influences the control strategy. And if controlled beliefs C_i are fixed, the maximization of $\overline{{}^2c_{p_i}}$ requires the selection of control group $\mathbf{C}$ to include nodes with highest $d(1+\gamma)^\alpha$ value in the network G [10].

The calculation of converged belief in Theorem II.3 requires the information of the degree list, clustering coefficient list and clustering weight α of network G. The clustering weight α is obtained by a learning process, which will be introduced in Section III. When calculating the expected control power, the information needed by the generalized Markov Graph model is still far less than the information of adjacency matrix A. And due to the fact that η could be rewritten as [10]:

$$\eta = (\Sigma_{i=1}^N d_i(1+\gamma_i)^\alpha)^2 - \Sigma_{i=1}^N d_i(1+\gamma_i)^\alpha,$$

the computational cost of such calculation is $O(N)$, which is the same as that in preferential attachment model.

III. Experiments on Converged Belief Estimation

The preferential attachment model and the generalized Markov Graph model are both tested on real network data [8]. There are 3 different types of social networks: on-line social networks, p2p transmission networks and physicist collaboration networks. Each of these three different social networks includes several subtypes of networks. On-line social networks include Slashdot network data of August 2008 and of February 2009, Wiki-vote network data and Epinions network data. P2p transmission networks include Gnutella network data at five different times. And physicist collaboration networks include collaboration networks of physicists studying astrophysics, condensed Matter Physics, theoretical high-energy Physics, experimental high-energy Physics and general relativity. The names of these 14 subtypes of networks will be denoted by indices: $1, 2, \ldots, 14$. From each of these 14 networks, we sampled 50 sub-networks using the same sampling method.

In this experiment, the accuracy of estimations of converged belief $\mathbf{B}(\infty)$ of both models is tested. The control strategy is set to push neutral public opinions towards positive opinions. The initial beliefs w_i of nodes in network are set to be neutral, i.e, they obey a uniform distribution on $[-1, 1]$. For both the preferential attachment model and the generalized Markov Graph model, 25 randomly chosen networks are selected from each sub-category of networks, and used as a testing set. The other 25 networks for each sub-category of networks are used as a training set for the generalized Markov Graph model to learn a clustering weight α.

A. *Preferential Attachment Model*

For each of the network samples in the testing set of each sub-category of networks, a degree list is recorded. The control set $\mathbf{C}$ is set to be the top 5% nodes with the highest degree in the network, and thus the number of control nodes is set to be: $c = \lceil 5\% N \rceil$. The controlled belief C_i for nodes in the control set are set to be 1. For each network sample, the private belief list w_i is generated 100 times according to a uniform distribution on $[-1, 1]$.

For each of the generated private belief lists, the expected value of control power $\overline{{}^{1}c_{p_i}}$ is calculated as shown in Section II-A2 for each of the generated private lists.

And the same network is used for calculating the exact value of control power according to Eq.(3). Then the relative error between the $\overline{{}^{1}c_{p_i}}$ and the exact value is recorded as the relative error for this network sample under this belief list. The 100 relative errors for all generated private belief list are then averaged and recorded as the relative error of this network sample. Next, the relative errors for all network samples in the testing set of each sub-category of networks are averaged and recorded as the relative error of preferential attachment model on this sub-category. The relative errors for 14 different sub-categories of networks are shown in Fig.(2) to compare with the results from the generalized Markov Graph model.

B. *Generalized Markov Graph Model*

The first step in a Generalized Markov Graph model is to learn the clustering weight α from the training sets for each sub-category of networks. For each training network sample, the degree list and clustering coefficient list are recorded. The control set $\mathbf{C}$ is set to be the top 5% nodes with highest degree in the network, and thus the number of control nodes is set to be: $c = \lceil 5\% N \rceil$. The controlled belief C_i for nodes in the control set are set to be 1. And 100 private belief lists sampled from a uniform distribution on $[-1, 1]$ is prepared.

In the training set of each sub-category of networks, for each network sample and an arbitrary value of α, the expected control power $\overline{{}^{2}c_{p_i}}$ is calculated for each of the 100 private beliefs. The exact solution of control power is also calculated for the 100 generated private belief lists. Then the relative error between the $\overline{{}^{2}c_{p_i}}$ and the exact value of the same private belief list is calculated, then averaged across all generated private belief lists, and recorded as the relative error for this network sample. For all 25 network samples in the training set, the relative errors are calculated and their average value is recorded as the relative error of this sub-category of networks. The clustering weight α for the 14 subtypes of networks are: [-1.00, -1.00, -0.50, -0.30, -0.20, 1.29, 1.52, 1.45, 1.12, 1.48, -2.57, -2.57, -2.84, 0.52].

The learnt α is then used to test the performance of the generalized Markov Graph model on the testing set. For each of the network samples in the testing set, the degree list and the clustering coefficient list are recorded. The control set is set as the same as in the training set. The later process is the same as that in the Preferential Attachment model. The relative errors for 14 different sub-categories of networks are shown in Fig.(2), together with the result from the preferential attachment model.

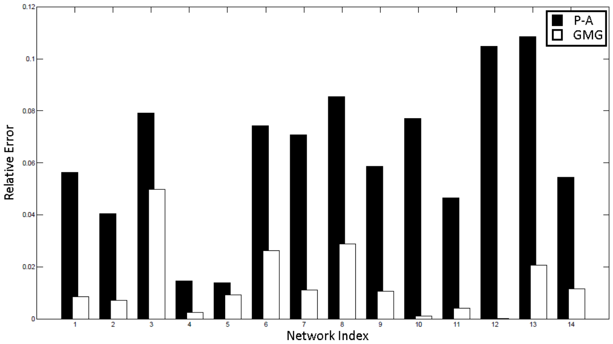

Fig. 2: Relative error of control powers for the preferential attachment model (P-A) and the generalized Markov Graph model (GMG) .

IV. Conclusion and Future Research

In this paper, we introduced a information flow model combined with two network models to simulate and calculate the converged beliefs of agents, as well as optimize the control strategy in a social network. Compared to a direct calculation of the converged beliefs, these two models use less information and require less computational power, but still with a good accuracy. In addition, the Generalized Markov Model outperforms the preferential attachment model since it has a more realistic assumption and uses more information.

References

[1] Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *Review of Economic Studies, Oxford University Press*, 78(4):1201–1236, 2008.

[2] Réka Albert and Albert Barabási. Statistical mechanics of complex networks. *REVIEWS OF MODERN PHYSICS*, 74:47–96, JANUARY 2002.

[3] Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The Review of Economic Studies, Vol. 65, No. 3 (Jul., 1998), pp. 595-621.*

[4] Abhijit V. Banerjee. A simple model of herd behavior. *THE QUARTERLY JOURNAL OF ECONOMICS*, CVII, Issue 3, August 1992.

[5] Douglas Gale and Shachar Kariv. Bayesian learning in social networks. *Games and Economic Behavior*, 45:329–346, November 2003.

[6] Usman A. Khan, Soummya Kar, and Jos M. F. Moura. Higher dimensional consensus: Learning in large-scale networks. *IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 58, NO. 5, MAY 2010.*

[7] StanFord University. Stanford large network dataset collection, retrieved from: http://snap.stanford.edu/data/, Sep 2011.

[8] Tian Wang, H. Krim, and Y. Viniotis. A generalized markov graph model: Application to social network analysis. *IEEE Journal of Selected Topics in Signal Processing*, 7:318,332, April 2013.

[9] Tian Wang, Hamid Krim, and Yannis Viniotis. Analysis and control of beliefs in social networks (in preparation).