Model Selection Of Graph Signage Models Using Maximum Likelihood (Student Abstract)

Angelina Brilliantova, Ivona Bezáková

Rochester Institute of Technology, 102 Lomb Memorial Drive, Rochester, NY 14623
lb9849@rit.edu

Abstract

Complex systems across various domains can be naturally modeled as signed networks with positive and negative edges. In this work, we design a new class of signage models and show how to select the model parameters that best fit real-world datasets using maximum likelihood.

Introduction

Networks with positive and negative edges (*signed networks*) are ubiquitous across various domains, including sociology, epidemiology, and gene regulation. Accounting for edge types helps substantially with many important network problems, such as link prediction, node ranking, and network synchronization (Li, Fang, and Zhang 2017).

Multiple frameworks for solving problems on signed networks exist – but, to be successful in real-world applications, their performance needs to be benchmarked on realistic synthetic datasets, which remain scarce for signed networks. Moreover, the difference between network structures from different domains might affect the assessment of algorithms' applicability and performance. Therefore, there is a need for generative domain-specific models to create new observations from the probability distribution underlying real-world instances and help algorithms avoid overfitting.

Many generative models exist for unsigned networks, but few generative models for signed networks, e.g. (Derr, Aggarwal, and Tang 2018). For the unsigned networks, model selection frameworks assessing the model fit to the dataset based on the maximum likelihood have been developed and successfully applied (Bezáková, Kalai, and Santhanam 2006). Such frameworks find the model and its parameters with the highest probability of generating the observed network and allow for a more rigorous comparison of model fit compared to analyzing high-level network characteristics such as degree distributions. We extend this approach to model selection for network signages.

In this work we design a new class of signage models and show how to select the model and its parameters that best fit the real-world datasets, using maximum likelihood. Unlike in the works cited above, our signage models work for scenarios in which the graph topology is formed first and later refined with the signs for the edges based on the nodes' attributes. A real-life example is forming social interactions in a closed community (e.g., dorms) in which you first get acquainted with the people you live and interact with and later decide on your attitude to them.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Graph Signage Model

A *signed* graph $G^{\pm} = (V, E, \mathcal{A})$ consists of a directed graph $G = (V, E)$ with vertex set V of size n and edge set E, and a signage function $\mathcal{A} : E \to \{+, -\}$.

We propose a graph signage model, where the signs of the edges of a given directed graph $G = (V, E)$ are driven by a random latent node partition C and a parameter matrix ξ. Let $S = \{A, R\}$ be a set of symbolic node labels, q be a probability, and ξ be an $|S| \times |S|$ matrix of probabilities $\xi_{x,y}$, where $x, y \in S$. The signage model first generates an independent random label for each vertex, A with probability q and R otherwise, creating a random latent node partition $C : V \to S$. Then, for each edge $(u, v) \in E$, the model assigns the sign $+$ to this edge with probability $\xi_{C(u),C(v)}$, and the sign $-$ otherwise, obtaining a signed graph $G^{\pm}$. The parameters of the model are combined in the tuple $\Theta = (\xi, q)$.

Model Selection

Given a signed graph $G^{\pm} = (V, E, \mathcal{A})$, we are looking for such a combination of parameters Θ that has the highest probability $P(\mathcal{A}|\Theta)$ of generating the edge signs $\mathcal{A}$. For a fixed node partition C, we can find the corresponding probability of the edge signs as:

$$P(\mathcal{A}|\Theta, C) = \prod_{(u,v)\in E} \mathcal{P}^{\mathcal{A}(u,v)}_{C(u),C(v)}, \tag{1}$$

where, for $s_1, s_2 \in S$, let $\mathcal{P}^{+}_{s_1,s_2} := \xi_{s_1,s_2}$ and $\mathcal{P}^{-}_{s_1,s_2} := 1 - \xi_{s_1,s_2}$.

In our model the labels of vertices are unknown, so the probability (likelihood) of generating edge signs $\mathcal{A}$ is a sum of conditional probabilities over all possible node partitions: $\mathcal{L}(\Theta) := P(\mathcal{A}|\Theta) = \sum_{C\in\mathcal{C}} P(\mathcal{A}|\Theta, C) \cdot P(C)$. Our framework aims to find model parameters Θ_{MLE} with the highest probability (or likelihood) of realizing the edge signs $\mathcal{A}$: $\Theta_{\mathrm{MLE}} = \arg\max_{\Theta} \mathcal{L}(\Theta)$.

We express $\mathcal{L}(\Theta)$ as the product of ratios of pairs of likelihoods on decreasingly smaller spaces of node partitions:

the reduction in space is achieved by fixing the labels for some nodes. The computation is based on the problem's self-reducibility as described below.

Likelihood estimation through the product of ratios Let $V = \{v_1, v_2, \ldots, v_n\}$. For $s_1, \ldots, s_j \in S$, we define $\mathcal{C}_j^{[s_1,\ldots,s_j]}$ as the set of node partitions $C : V \to \{A, R\}$ such that $C(v_i) = s_i$ for every $i \leq j$ (vertices $v_1, \ldots, v_j$ have their labels determined by $s_1, \ldots, s_j$). Notice that $\mathcal{C}_0$ is the set of all node partitions (with no restrictions) and that $|\mathcal{C}_j^{[s_1,\ldots,s_j]}| = 2^{n-j}$ for any $s_1, \ldots, s_j$.

For our self-reducibility approach, let us fix a "master" node partition $\tilde{C} : V \to \{A, R\}$ that gradually more and more vertices will adhere to. Let $Z_j := \sum_{C_j \in \mathcal{C}_j^{[\tilde{C}(v_1),\ldots,\tilde{C}(v_j)]}} P(\mathcal{A}|\Theta, C_j) \cdot P(C_j|j)$ where $P(C_j|j)$ is the probability of C_j restricted to vertices $v_{j+1}, \ldots, v_n$. Notice that $P(C_j|j)$ is a probability distribution over the node partition subspace $\mathcal{C}_j^{[\tilde{C}(v_1),\ldots,\tilde{C}(v_j)]}$, and, therefore, Z_j is the likelihood of Θ restricted to this subspace. We estimate $\mathcal{L}(\Theta) = Z_0$ via the following product of likelihood ratios Z_{j+1}/Z_j:

$$P(\mathcal{A}|\Theta) = Z_0 = \frac{Z_n}{\prod_{j=0}^{n-1} Z_{j+1}/Z_j}, \tag{2}$$

where $Z_n = P(\mathcal{A}|\Theta, \tilde{C})$ can be easily computed by (1).

The main computational difficulty arises from the fact that the calculation of Z_{j+1}/Z_j might depend on the labels of the remaining $n - j$ nodes, which, at that point, are still unassigned in the course of the algorithm. Instead of exact computation, we estimate each ratio via a Markov Chain Monte-Carlo (MCMC) sampling of node partitions on the subspace corresponding to Z_j, with first j vertex labels assigned.

MCMC sampling to estimate the ratios Suppose $\tilde{C}(v_1), \ldots, \tilde{C}(v_{j+1})$ have been already defined. Let us define $w_j(C_j) := P(\mathcal{A}|\Theta, C_j) \cdot P(C_j|j)$ as the weight of the node partition $C_j \in \mathcal{C}_j^{[\tilde{C}(v_1),\ldots,\tilde{C}(v_j)]}$. Notice that $Z_j = \sum_{C_j \in \mathcal{C}_j^{[\tilde{C}(v_1),\ldots,\tilde{C}(v_j)]}} w_j(C_j)$. We will estimate Z_{j+1}/Z_j by randomly generating C_j with probability proportional to its weight. Therefore, C_j will be generated with probability $\mu_j(C_j) := w_j(C_j)/Z_j$. We generate samples C_j from the μ_j distribution and compute the proportion of samples with $C_j(v_{j+1}) = \tilde{C}(v_{j+1})$. As the number of samples increases, this quantity converges to the ratio Z_{j+1}/Z_j (times an easy-to-compute factor). See Algorithm 1. For numerical stability reasons, we do not choose $\tilde{C}$ upfront but generate it during the algorithm, see lines 10-13.

We use the Metropolis-Hasting technique to design a Markov chain with the desired stationary distribution μ_j: The state space is $\Omega_j := \mathcal{C}_j^{[\tilde{C}(v_1),\ldots,\tilde{C}(v_j)]}$ and the Markov chain transitions are defined as follows. Let C_j be the current state. Repeat T times:

- Choose a uniformly random $z \in \{j+1, \ldots, n\}$. Let C'_j be identical to C_j, except $C'_j(v_z) = \bar{C}_j(v_z)$ (the label of v_z is opposite in C_j and C'_j).
- With probability $\min\{1, \frac{w_j(C'_j)}{w_j(C_j)}\}$ move to state C'_j. Otherwise, stay at C_j.

Algorithm 1: ESTIMATE-LIKELIHOOD($G^{\pm}$, Θ, K, T)

1: Throughout the algorithm, $\tilde{C} : \{v_1, \ldots, v_j\} \to \{A, R\}$.
2: **for** $j = 0$ to $n - 1$ **do**
3: Let $\text{sum}_A = 0$ and $\text{sum}_R = 0$.
4: **for** $k = 1$ to K **do**
5: C = SAMPLE($\mathcal{A}, \Theta, \tilde{C}, T$) (use Markov chain)
6: **if** $C(v_{j+1}) = A$ **then** Increment sum_A
7: **else** Increment sum_R
8: Let $S_k^A = \text{sum}_A/k$ and $S_k^R = \text{sum}_R/k$
9: **end for**
10: **if** $S_k^A \geq \frac{1}{2}$ **then**
11: Let $\tilde{C}(v_{j+1}) = A$ and $S_j = S_k^A/q$
12: **else**
13: Let $\tilde{C}(v_{j+1}) = R$ and $S_j = S_k^R/(1-q)$
14: **end if**
15: **end for**
16: Use (1) to compute Z_n
17: **return** $\frac{Z_n}{\prod_{j=1}^{n} S_j}$

Experiments

We evaluated our algorithm on signed networks of gene regulatory interactions of two bacterial species - *E.coli* ($n = 1922$) and *B. subtilis* ($n = 2563$) – obtained from public databases. Algorithm 1 estimates the likelihood of a single Θ instance. To find Θ_{MLE} with the best likelihood, we searched over the 5-dimensional Θ-space, divided into a grid with 3 values per each dimension. We refined the search by exploring the neighborhood of the best Θ-candidate obtained in the previous step. In total, we tested 675 Θ instances per dataset using parallel implementation on our university's computing cluster. The likelihood of a single Θ instance was estimated on a single core and took up on average 12 hours. A core has Intel®Xeon®Gold 6150 CPU @ 2.70GHz.

Acknowledgements

We thank the NSF (awards DUE-1821459 and 1819546) and RIT Research Computing for their support.

References

Bezáková, I.; Kalai, A.; and Santhanam, R. 2006. Graph model selection using maximum likelihood. In *Proceedings of the 23rd International Conference on Machine learning*, 105–112.

Derr, T.; Aggarwal, C.; and Tang, J. 2018. Signed network modeling based on structural balance theory. In *Proceedings of the ACM international conference on information and knowledge management*, 557–566.

Li, X.; Fang, H.; and Zhang, J. 2017. Rethinking the link prediction problem in signed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4955–4956.