Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

On the Interpretability of Thresholded Social Networks

Oren Tsur[†‡§] orentsur@seas.harvard.edu
David Lazer[†‡] d.lazer@neu.edu
[†]Network Science Institute, Northeastern University
[‡]Institute for Quantitative Social Science, Harvard University
[§]School of Engineering and Applied Sciences, Harvard University

Abstract

Understanding the factors of network formation is a fundamental aspect in the study of social dynamics. Online activity provides us with abundance of data that allows us to reconstruct and study social networks. Statistical inference methods are often used to study network formation. Ideally, statistical inference allows the researcher to study the significance of specific factors to the network formation. One popular framework is known as Exponential Random Graph Models (ERGM) which provides principled and statistically sound interpretation of an observed network structure. Networks, however, are not always given set in stone. Often times, the network is "reconstructed" by applying some thresholds on the observed data/signals. We show that subtle changes in the thresholding have significant effects on the ERGM results, casting doubts on the interpretability of the model. In this work we present a case study in which different thresholding techniques yield radically different results that lead to contrastive interpretations. Consequently, we revisit the applicability of ERGM to thresholded networks.

1 Introduction

Online activity provides us with abundance of data, allowing us to reconstruct social networks and analyze the social processes and dynamics. Recent studies model complex contagion and information diffusion (Leskovec, Backstrom, and Kleinberg 2009; Yang and Leskovec 2010; Romero, Meeder, and Kleinberg 2011; Tsur and Rappoport 2015) as well as social polarization (Adamic and Glance 2005; Guerra et al. 2013; Andris et al. 2015). These works assume a given (mostly fixed) network and study its topology and its nodal activity but do not address the *formation* of the observed network. The social factors that drive the formation of a social network are of great interest for sociologists, political scientists and computer scientists alike. Exponential Random Graph Models (ERGM) is one popular framework for studying network formation[1] (Morris, Handcock, and Hunter 2008).

Ideally, the ERGM framework allow the researcher to study the significance of the contribution of specific factors to the network formation. Given a network and complex hypothesis, ERGM assigns significance values to the various terms in the hypothesis. ERGM, in that sense, provides principled and statistically sound interpretation of an observed network structure.

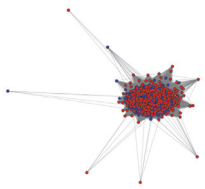
Networks, however, are not always given set in stone. Often times, the network is reconstructed by applying some thresholds on raw data. For example, consider a network based on online communication. A plausible approach is to add an edge (u, v) only if u refers[2] to v more than δ times, where the threshold δ is chosen arbitrarily in a heuristic manner. This thresholding (or edge pruning) is usually required for noise reduction – avoiding a "hairball" network and allowing a binary[3] representation of the network for efficient computation of statistical dependencies. Figure 1 illustrates some of the benefits of simple thresholding: removing noisy communications by applying a threshold of $\delta = 3$ transforms the network of Members of the U.S. Congress from a "hairball" mess to obvious partisan modules. This popular thresholding approach is commonly applied in the study of online social networks; two notable and influential examples are (Adamic and Glance 2005) and (Romero, Meeder, and Kleinberg 2011).

In this paper we show that subtle changes in the thresholding have significant effects on the ERGM results, casting doubts on the interpretability of the model. Specifically, we reconstruct the political network based on the communication of members of the U.S. Congress. We specify a social-science informed ERGM model and show how different thresholds, result in very different models, giving rise to very different and sometimes contradicting interpretations of the process governing the network formation. These results raise some concerns regarding the applicability of ERGM and similar approaches to thresholded networks.

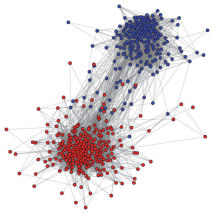
[1]Other statistical frameworks, similar in varying degrees are Latent Space Model (Hoff, Raftery, and Handcock 2002) and Quadratic Assignment Procedure (Krackhardt 1988).

[2]The type of reference/signal to apply a threshold on depends on the platform and the network semantics of interest (e.g. mentioning or retweeting on Twitter, liking on Facebook, linking in blogs or citing in academic publications).

[3]Some differences between binary and weighted networks are explored in (Barrat et al. 2004) among others.



(a) No thresholding.



(b) With thresholding $\delta = 3$.

Figure 1: Networks of Twitter communication of members of the U.S. congress. Color represent partisan affiliation. Both networks were reconstructed from the data described in Section 2 and were plotted using the exact same layout (Fruchterman-Reingold) and parameters.

2 Data

We consider the political network observed by curating the Twitter communication of members of the U.S. Congress and Senate (MoCS). We focus on this network for the following reasons: (i) We want a complex network of considerable yet manageable size, (ii) We ultimately care about interpretation thus we need a network of known actors allowing comparison to well developed theories from political science.

We collected all tweets posted by members of the U.S. Congress and Senate during a year spanning from April 2015 to April 2016, a total of about 400,000 tweets (posted by 520 active users). It is a common practice to look at user mentions (@) in reconstructing Twitter networks (Romero, Meeder, and Kleinberg 2011) or retweets (Guerra et al. 2013), among others. Mentions and retweets are seen as more accurate than just following relations as they demonstrate an active signal. It is common to assume that while a single mention can be anecdotal, reoccurring mentions of *v* by *u* provide a strong relational signal. We consider the following thresholds: 3,4,5 and 25, all appear in the literature. While it is clear that mentioning someone dozens of times is semantically different than mentioning her only a handful of times, choosing a threshold between 3 and 5 seems inconsequential. Indeed, we show that some standard network statistics remain stable given small variations in the thresholding (see section 4.1). However, as demonstrated in Section 4.2, these seemingly inconsequential variations yield major differences in the network semantics as modeled by ERGM.

3 Exponential Random Graph Models

While we can use various statistics (e.g. centrality, modularity, degree distribution) to describe an observed network, these structural features do not provide much insight about the (social) processes governing the formation of the network. Generally speaking, the network at hand is one specific realization of a network instance out of a pool of possible networks. Some of these possible network may have the same exact topology (e.g. permutation of the nodes in the observed network) and some other networks may have a different structure. In order to understand the formation of the specific network we would like to allow statistical inference of the factors that may be involved in the network formation. The main challenge, however, is that standard statistical models assume uncorrelated variables, while network variables are inherently dependent.

There is a number of frameworks that support statistical inference on networks (see footnote 1). One approach that is gaining popularity among statisticians, social and political scientists is ERGM – Exponential Random Graph Models. A detailed description of the ERGM framwork and its evolution can be found in (Snijders 2002; Morris, Handcock, and Hunter 2008; Schweinberger and Handcock 2015; Wilson et al. 2017). In this section we briefly describe the ERGM framework and specify our model.

The abstract form of an ERGM model is:

$$Pr(Y = y) = (\frac{1}{k})exp\sum_{A}\theta_A g_A(y) \quad (1)$$

Which reads as 'the probability of a specific realization y is a function of a set of nodal and/or dyadic features denoted by A, where each node or set of nodes either holds the feature or not'. Gibbs sampling is used in order to infer θ – the feature coefficients. A normalization factor k is used to enforce the learned model to be a probability function.

3.1 Naive Model - Number of Edges

The baseline model is the Erdős Rényi network (Erdös and Rényi 1959). Given a set of nodes and a general edge likelihood θ, the Erdős Rényi random network is generated by applying this likelihood at every possible pair of nodes. In our case, we observe a fixed number of edges and infer θ. We use this term as a baseline model since no nodal or dyadic terms such as homophily (nodal feature) or reciprocity (dyadic feature) are considered. Formally this term is defined as:

$$\sum_{i,j}\theta y_{ij}$$

Where i and j are nodes and $y_{ij} = 1$ iff the edge (i, j) is observed in the network. In the remainder of this section we further introduce a number of nodal and dyadic terms, motivate the use of these terms in modeling our network and define the formal ERGM model we use.

3.2 Nodal and Dyadic Terms

Reciprocity Reciprocity is one of the basic dyadic terms used in network analysis. The idea of quid pro quod is an inherent characteristic of many social interactions and can

be significant in network formation. The likelihood of v following u in a social network like Twitter may increased if we know that u follows v. Formally, we define the term as follows:

$$\sum_{ij} \theta_r y_{ij} y_{ji}$$

Where i and j ($i \neq j$) are nodes in the network and $y_{ij} = 1$ iff i follows j. This term is nonzero iff i follows j *and* j follows i.

Partisan Homophily We assume that nodes belonging to the same party are more likely to be connected. Formally this term is represented as:

$$\sum_{ij|p(i)=p(j)} \theta_p y_{ij}$$

Where $p(i)$ is the party of node i.

Leadership We assume that leadership positions (party leaders, speakers, whips) play a role in network formation. In the Twitter context, they may be more likely to be followed, replied to or retweeted. We denote the set of MoCs in leadership position as L and formally define the leadership term as:

$$\sum_{ij|,j\in L} \theta_L y_{ij}$$

Seniority We assume that seniority – the number of terms a MoC served – may effect the network formation in a numebr of possible ways – e.g. senior members with a longer shared history will entertain a more cliquey behavior or that senior MoCs may have a more central role in the network due to their experience. Formally we define seniority in a similar way to *leadership*:

$$\sum_{ij|t(j)>3} \theta_s y_{ij}$$

Where $t(j)$ is the number of terms[4] j served in the Congress/Senate.

Twitting Rate Since our network is based on (@) mentions, the number edges (incoming and outgoing) a node has may depend on the number of tweets she posts. High volume of tweets create visibility and may attract mentions by other nodes. Likewise, the likelihood of mentioning another user may increase simply as a function of the number of tweets a node posts. This term, therefore, serves as a nodal control term in a similar way to the way the Erdős Rényi term serves as a control based on the number of observed edges. The rate term is defined as:

$$\sum_{ij|tr(j)\in P_k} \theta_{tr} y_{ij}$$

Where $tr(j)$ is j's tweeting rate and P_k is the k's percentile of volume of tweets by user (node).

[4]We note the ambiguity of the term *term*, here used for the number of times j was elected for Congress.

3.3 Joint Network Model

We combine all the terms above to one model, providing a relatively simple yet seemingly powerful and well motivated model accounting to major factors that may govern the process of network formation. Using the notation in Equation 1, we use Gibbs sampling to jointly infer the θs in the following model:

$$Pr(Y = y) = (\frac{1}{k}) exp[\sum_{i,j} \theta y_{ij} + \sum_{ij} \theta_r y_{ij} y_{ji} + \sum_{ij|p(i)==p(j)} \theta_p y_{ij} + \sum_{ij|,j\in L} \theta_L y_{ij} + \sum_{ij|t(j)>3} \theta_s y_{ij} + \sum_{ij|tr(j)\in P_{\{1,2,3\}}} \theta_{tr} y_{ij}] \quad (2)$$

4 Results

In this work we aim to examine two main assumptions that are typically held in network analysis. We show how their in/validity effects the social interpretation of the network formation based on statistical inference. The two assumptions are:

1. Small thresholding values are used to remove noisy ("occasional") interaction, hence small variance in thresholding (low) values has minimal effect on network statistics.
2. Different thresholding in network representation captures different social/network semantics (e.g. high threshold is expected to only capture the strongest ties). Therefore, high variance in the thresholding values results in significant difference in network statistics.

In Section 4.1 we show that both assumptions hold for two standard network statistics: modularity and betweenness centrality. In Section 4.2 we show that while the second assumption holds, the first, seemingly straight forward, assumption is not supported by the statistical model. Violation of these assumptions leads to major concerns regarding the common (and naive) application of the ERGM framework on thresholded networks.

4.1 General Network Statistics

Results for various thresholds are presented in Table 1. The modularity values for $\delta \in \{3, 4, 5\}$ change only slightly from 0.317 to 0.328. Similarly, average betweenness centrality ranges from 0.0033 to 0.0036. The insignificant change in the values of both modularity and centrality validates the first assumption.

Comparing the modularity and betweenness values of $\delta \in \{3, 4, 5\}$ to $\delta = 25$ we see a 30% increase in modularity while the average betweenness centrality drops significantly. These results support the second assumption and are in line with the literature, e.g. (Adamic and Glance 2005).

4.2 Instability of Statistical Inference

Using ERGM to infer the θ coefficients for the terms of interest tells a very different story. Table 2 presents the significance levels of the different terms in the model. The ob-

	$\delta = 3$	$\delta = 4$	$\delta = 5$	$\delta = 25$
Modularity	0.3177	0.3230	0.3280	0.4354
Betweenness	0.0033	0.0036	0.0035	0.00005

Table 1: Modularity and average betweeness centrality values given different thresholding (δ).

Term	$\delta = 3$	$\delta = 4$	$\delta = 5$	$\delta = 25$
#Edges	***	***	***	***
Reciprocity	***	***	***	—
Partisan homophily	*(R)	—	—	—
Leadership position	**(D)	* (D)	.(D)	**
Seniority	*** (↑)	**(↑)	**(↓)	
Tweeting rate	—	—	* (↓)	—

Table 2: Significance of various terms obtained by ERGM and given different thresholding (δ). ↑(↓): significance in higher (lower) values of the term. — for insignificance.

served number of edges is significant for all δ values (as expected). Reciprocity is significant for $\delta \in \{3, 4, 5\}$ and insignificant for the higher value of δ (suggesting other terms come into play for this nonreciprocal network). While the given number of (observed) edges and reciprocity are stable across the low δ values, the other terms lend themselves to inconsistent interpretation. Partisan homophily is significant only for $\delta = 3$ and only for Republicans. Leadership position is significant, though diminishing, only for Democrats for $\delta \in \{3, 4, 5\}$ and for both parties for $\delta = 25$. Seniority matters only for $\delta \in \{3, 4, 5\}$, however it seems that if we threshold with $\delta \in \{3, 4\}$ it is the senior members who attract incoming edges (due to their seniority?) while using a threshold of $\delta = 5$ shows that the younger MoCs attract incoming edges, suggesting a very different interpretation (they are younger and more active? they show higher tendency to reciprocate? they are more cliquey?). These results suggest the first assumption is implausible, an alarming message to the researcher who is interested in the social interplay between the nodes in the network, suggesting that "noise reduction" and "binarization" should be performed in a more informed manner.

5 Conclusion

We showed that the common practice of thresholding for network pruning can lead to instability in results of statistical inference, and thus to wrong or incoherent interpretation of the social dynamics that shaped the observed network. While we cannot yet offer a proven remedy for this problem, we wish to conclude by mentioning two directions that may help in addressing this issue: (i) prune the network in a statistically informed ways e.g. (Serrano, Boguná, and Vespignani 2009; Radicchi, Ramasco, and Fortunato 2011; Dianati 2016), and (ii) apply the newly developed Generalized-ERGM (GERGM) framework (Wilson et al. 2017) which theoretically allows inference with weighted edges.

Acknowledgments This work was generously supported by NSF CDI Type II grant 501960.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, 36–43. ACM.

Andris, C.; Lee, D.; Hamilton, M. J.; Martino, M.; Gunning, C. E.; and Selden, J. A. 2015. The rise of partisanship and super-cooperators in the us house of representatives. *PloS one* 10(4):e0123507.

Barrat, A.; Barthelemy, M.; Pastor-Satorras, R.; and Vespignani, A. 2004. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* 101(11):3747–3752.

Dianati, N. 2016. Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Physical Review E* 93(1):012304.

Erdös, P., and Rényi, A. 1959. On random graphs, i. *Publicationes Mathematicae (Debrecen)* 6:290–297.

Guerra, P. H. C.; Meira Jr, W.; Cardie, C.; and Kleinberg, R. 2013. A measure of polarization on social media networks based on community boundaries. In *ICWSM*.

Hoff, P. D.; Raftery, A. E.; and Handcock, M. S. 2002. Latent space approaches to social network analysis. *Journal of the american Statistical association* 97(460):1090–1098.

Krackhardt, D. 1988. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks* 10(4):359–381.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 497–506. Citeseer.

Morris, M.; Handcock, M. S.; and Hunter, D. R. 2008. Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software* 24(4).

Radicchi, F.; Ramasco, J. J.; and Fortunato, S. 2011. Information filtering in complex weighted networks. *Physical Review E* 83(4):046101.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM.

Schweinberger, M., and Handcock, M. S. 2015. Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(3):647–676.

Serrano, M. Á.; Boguná, M.; and Vespignani, A. 2009. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences* 106(16):6483–6488.

Snijders, T. A. 2002. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure* 3(2).

Tsur, O., and Rappoport, A. 2015. Dont let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes. *Proc. ICWSM*.

Wilson, J. D.; Denny, M. J.; Bhamidi, S.; Cranmer, S.; and Desmarais, B. 2017. Stochastic weighted graphs: Flexible model specification and simulation. *Social Networks* 49:3747.

Yang, J., and Leskovec, J. 2010. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, 599–608. IEEE.