Measuring the Statistical Significance of Local Connections in Directed Networks

James D. Wilson, Shankar Bhamidi, and Andrew B. Nobel
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
{jameswd, bhamidi, nobel}@email.unc.edu

Abstract

Partitioning a network into different communities so that vertices of the same community share meaningful density- and pattern-based similarities is an important area of research in the field of network science. For directed networks identifying communities turns out to be especially challenging since the directed nature of the edges makes it difficult to evaluate and interpret the significance of a candidate community. In this paper, we consider the strength of connections from a single vertex to a prespecified collection of vertices in directed networks. We propose a methodology to measure the statistical significance of these connections through the use of p-values derived from a directed configuration null model. We derive the asymptotic distribution of the number of edges between a vertex and a community under the null model and show how to calculate p-values using this reference distribution. Using both simulated and real data sets we show that these conditionally based p-values can provide novel insights into the local structure of directed networks.

1 Introduction

Networks arise in the modeling and understanding of a host of complex systems, ranging from biological networks such as protein and gene interaction networks [2, 13], social networks modeling collaborations, friendships, and other ties between individuals [8, 9, 21], and the world wide web, namely the hyperlink structure between webpages or webblogs [1, 15]. In many such systems agents influence each other in an asymmetric way and the associated network is directed. This includes gene-interaction networks, where the expression of one gene causes or suppresses the expression of another, and webblog networks where one blog references another by posting a hyperlink. Abstractly such networks are represented via graphs where every edge has a specified direction.

Empirically, networks have the tendency to cluster into communities. In undirected networks, a community is informally said to be a collection of vertices that share more edges within their own community than they share with vertices outside the community. In the context of undirected networks, there is an enormous amount of literature and a wide array of algorithmic techniques to extract such structures from data, see e.g. the surveys [7, 17] and the references therein. There has been some recent work on the community structure in directed networks including a directed version of modularity [3, 12], as well as a directed variant of spectral clustering [27]. Alternatively, the community structure of a directed network can be estimated through fitting a directed stochastic block model [2, 19, 22]. See [14] for a recent review of community detection in directed networks.

Though much work has been done on identifying community structure in both undirected and directed networks, quantifying the significance of such local structures has been much less explored. In undirected networks, the focus is typically on the significance of a proposed partition of the network [5, 20, 25], though some authors have focused on more granular features of the network

[11, 23]. To the best of our knowledge, there has been no prior work on assessing the statistical significance of local structures in directed networks.

The main aim of this work is to quantify the statistical significance of connections from a vertex to a collection of vertices in a directed network. That is, given a collection of vertices B and a single vertex u, either within or outside of B, we aim to measure the strength of connection from u to B. To fix a concrete example, consider the Adamic and Glance dataset on political blogs [1] shown in Figure 1. One may be interested in knowing the association of blogs to a collection of blogs B, where the blogs of B all have a similar feature such as political affiliation or posting habits, or B may be a community detected from applying a detection method on the network.

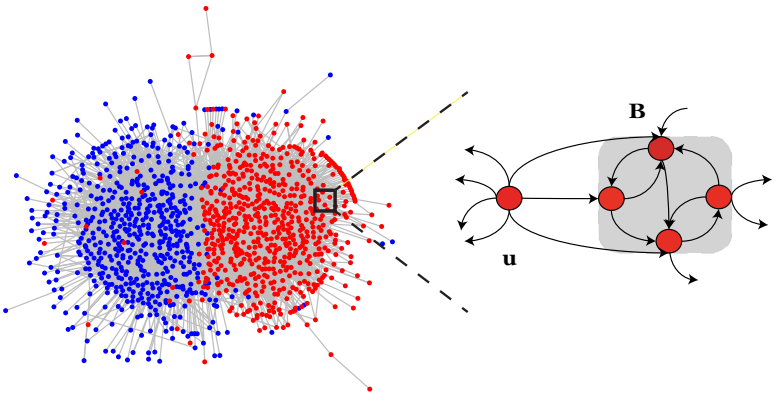

Figure 1: An example of analyzing local connections within the political blog network where directed edges represent posted hyperlinks between blogs. In this example, we consider measuring the strength of connections between a particular webblog u and a collection of webblogs B based on the hyperlink structure of these blogs.

We measure the strength of affiliation between a vertex and a collection of vertices by contrasting the observed number of edges between the two groups with what is expected under a random network with no preferential attachment, namely, the directed configuration model. Under this model, we show that the (random) number of connections between a vertex and a collection is approximately binomial. We use this approximate distribution as a reference measure to quantify the significance of observed connections through a p-value. Numerical simulations (Figure 2) suggest that this approximation is valid even for networks of moderate size ($n \geqslant 1000$). We demonstrate the effectiveness of our methodology on networks with pre-defined directed community structure, and through various simulated networks show how the p-values can identify significant local patterns in the network. We apply our proposed methodology to the political blog dataset of [1] and show that by quantifying local significance, one can detect interesting local features of this hyperlink structure of the network beyond what is available from standard community detection techniques.

2 Related Work

We briefly describe the related work done for networks in the undirected regime, as we know of no prior work on directed networks. The work closest to this paper is [23] where a null configuration model and the corresponding p-values were used to develop a testing-based procedure that extracts statistically significant communities. In [11], the statistical significance of a collection of vertices is assessed via measuring the probability of finding a subset with similar connectivity patterns in an (undirected) configuration model. The same authors later develop a community detection algorithm based on the order statistics of these probability scores in [10]. In [5] and [25], the significance of a community is assessed through comparison with the vertex features of the network. Here, the features of the network act as a ground-truth for which the significance of observed community structure can be assessed. The authors in [20] consider the statistical significance of a partition of an undirected network, where the modularity of a potential partition is first calculated and then

compared to the probability of observing this modularity of an equally sized partition in a random network.

3 Statistical Model and Framework

3.1 The Directed Configuration Model

Throughout we let $G_o = (V, E)$ be an observed directed network with n vertices, where G_o possibly contains self-loops or multiple edges. Assume that G_o has vertex set $V = [n] = \{1, \ldots, n\}$. The edge set E of G_o contains all (ordered) pairs (i, j) such that $i, j \in [n]$ and there is a directed link from i to j in G_o, with repetitions for multiple edges. For vertex $u \in [n]$ let $d_o^{in}(u)$ denote the in-degree and $d_o^{out}(u)$ denote its out-degree. Denote the in-degree sequence of G_o by $\mathbf{d_o^{in}} = \{d_o^{in}(1), \ldots, d_o^{in}(n)\}$ and the out-degree sequence by $\mathbf{d_o^{out}} = \{d_o^{out}(1), \ldots, d_o^{out}(n)\}$. Note that $\sum_u d_o^{in}(u) = \sum_u d_o^{out}(u) = |E|$.

Our analysis begins with a directed stochastic network model that is derived from the in- and out- degree sequences of G_o, specifically, the *directed configuration model*, which we denote by DCM($\{\mathbf{d_o^{in}}, \mathbf{d_o^{out}}\}$). The directed configuration model is a probability measure on the family of graphs with vertex set $[n]$, in-degree sequence $\mathbf{d_o^{in}}$, and out-degree sequence $\mathbf{d_o^{out}}$ that reflect a random assignment of directed edges. This model is a natural extension of the well known undirected configuration model [4, 6, 16] and has been used for modularity based community detection algorithms [3, 12]. An important characteristic of the directed configuration model is its ability to capture and preserve strongly heterogeneous degree distributions that are often encountered in real network data sets, as well as preserving the directed nature of the observed network.

The directed configuration model has a simple two stage generative form. First, each vertex $u \in [n]$ is assigned $d_o^{in}(u)$ inward pointing and $d_o^{out}(u)$ outward pointing directed half-edges. At the next stage, two half-edges - one inward and one outward - are chosen uniformly at random and connected to form a directed edge. These two half-edges are removed from the set of available half-edges. This procedure is repeated sequentially by picking at each stage a random inward pointing and outward pointing half-edge to connect until all half-edges are connected. We write $\hat{G} = ([n], \hat{E})$ to denote the random network generated by this procedure. Note that even if G_o is simple, $\hat{G}$ may contain self loops and multiple edges.

3.2 Asymptotic Results and Assessing the Significance of Local Connections

Under the DCM($\{\mathbf{d_o^{in}}, \mathbf{d_o^{out}}\}$) model there are no preferential attachments among vertices $[n]$. Thus, the model provides a reference measure against which the statistical significance of the connections from a vertex to a collection of vertices in G_o can be assessed. The more the observed number of directed edges from the expected number under the DCM, the greater the significance of the connection.

Let G_o be an observed network and $\hat{G}$ its associated random DCM network. Given a vertex $u \in [n]$ and a vertex set $B \subset [n]$ define

$$d_o(u : B) = \sum_{v \in B} \sum_{e \in E} \mathbb{I}(e = (u, v)) \tag{1}$$

to be the number of directed edges pointing *from* vertex u *to* some vertex in B in the observed network. Write $\hat{d}(u : B)$ for the random variable specifying the number of edges originating from u and ending in B in $\hat{G}$. Then $\hat{d}(u : B)$ takes values in the set $\{0, 1, \ldots, d^{out}(u)\}$. Note that $d_o(u : B) = \hat{d}(u : B) = d_o^{out}(u)$ when $B = [n]$.

Recall that the total variation distance between two probability mass functions $\mathbf{p} := \{p(i)\}_{i \geqslant 0}$ and $\mathbf{q} := \{q(i)\}_{i \geqslant 0}$ on the natural numbers $\mathbb{N}$ is defined by:

$$d_{TV}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \sum_{i=1}^{\infty} |p(i) - q(i)|$$

We now state a theorem that describes the approximate distribution of $\hat{d}(u : B)$ when the size of the network n is large.

Theorem 1. *Let* $\left\{\mathbf{d}_{o,n}^{in}, \mathbf{d}_{o,n}^{out}\right\}_{n\geqslant 1}$ *be the in- and out- degree sequences of an observed sequence of graphs* $\{G_o^n\}_{n\geqslant 1}$ *where* G_o^n *is a graph of size* n *and associated edgeset* E_n*. Let* $\{\hat{G}^n\}_{n\geqslant 1}$ *be the corresponding random graphs on* $[n]$ *constructed via the directed configuration model. Let* F_n *be the empirical distribution of* $\mathbf{d}_{o,n}^{in}$ *and* H_n *the empirical distribution of* $\mathbf{d}_{o,n}^{out}$*. Assume that there exist cumulative distribution functions* F *and* H *on* $[0,\infty)$ *with* $0 < \mu_1 := \int_{\mathbb{R}^+} x dF(x) < \infty$ *and* $0 < \mu_2 := \int_{\mathbb{R}^+} x dH(x) < \infty$ *such that*

$$F_n \xrightarrow{\text{w}} F, \qquad H_n \xrightarrow{\text{w}} H \tag{2}$$

and

$$\int_{\mathbb{R}^+} x\, dF_n(x) \;\to\; \mu_1, \qquad \int_{\mathbb{R}^+} x\, dH_n(x) \;\to\; \mu_2 \tag{3}$$

Fix $k^{out} \geqslant 1$ *and a vertex* $u = u_n \in [n]$ *with out-degree* $d_n^{out}(u) = k^{out}$*. Let* $B_n \subseteq [n]$*,* $n \geqslant 1$*, be a sequence of sets of vertices.*
Then, the random variable $\hat{d}_n(u : B_n)$ *is approximately Binomial*$(k^{out}, p_n(B_n))$ *in the sense that*

$$d_{TV}(\hat{d}_n(u : B_n), \textit{Bin}(k^{out}, p_n(B_n))) \to 0 \tag{4}$$

as $n \to \infty$*. Here,*

$$p_n(B_n) \;=\; \frac{\sum\limits_{v\in B_n} d_n^{in}(v)}{\sum\limits_{w\in[n]} d_n^{in}(w)} \;=\; \frac{1}{|E_n|}\sum_{v\in B_n} d_n^{in}(v) \tag{5}$$

A proof of this theorem is given in the Supplementary material. Theorem 1 gives us that under the directed configuration model, the number of directed edges from a vertex with out-degree k^{out} to a collection of vertices B is approximately binomial on $\{0, 1, \ldots, k^{out}\}$ with probability equal to the relative proportion of the total in-degree of B to the entire network. As the directed configuration model DCM($\{\mathbf{d_o^{in}}, \mathbf{d_o^{out}}\}$) does not contain preferential connections between vertices, one can assess the strength of connection between a vertex u and collection B in G_o by comparing the observed number of connections, $d_o(u : B)$, with the random variable $\hat{d}(u : B)$. By treating $d_o(u : B)$ as an observed value of a test statistic that is distributed as $\hat{d}(u : B)$ under the null network model DCM($\{\mathbf{d_o^{in}}, \mathbf{d_o^{out}}\}$), the probabilities

$$p(u : B) = P(\hat{d}(u : B) \geqslant d_o(u : B)) \tag{6}$$

have the form of a p-value for testing the hypothesis that u does not strongly link to B. Small values of $p(u : B)$ indicate that there are more edges from u to B than expected under the directed configuration model on the same vertex set. We use these p-values to quantify the strength of connection from any vertex to any fixed collection of vertices.

4 Numerical Study

4.1 Convergence Rate under the DCM

We first empirically investigate the convergence rate to zero of the total variation distance of $\hat{d}(u : B)$ from the binomial distribution as given in Equation (4). We construct directed configuration models $\mathcal{G}_n$ of size n where each vertex is first independently assigned an in-degree from a power law distribution with exponent $\tau = 3$. To ensure that the sum of the in- and out- degrees are equal, we randomly permute the in-degree sequence and assign each vertex an out-degree from this permuted sequence. We fix subsets $B_n \subseteq \mathcal{G}_n$ with $P(B_n) \approx 0.25$ for all n by letting B_n be a uniformly chosen random subset containing one fourth of the vertices in the network. We calculate the observed number of connections $d(u : B_n)$ for all vertices $u \in \mathcal{G}_n$ with out-degree k for a range of k between 3 and 10. We then calculate the total variation distance $d_{TV}(d(u : B_n), \text{Bin}(k, p(B_n)))$ using the empirical distribution of $d(u : B)$ for each fixed k. At each size n, we simulate 100 networks and calculate the total variation distance in this way for each network and each k. We

repeat this simulation across n from 500 to 10000 in increments of 500 and report the distribution of the total variation distance at each n for $k = 3$ in Figure 2. Even for n as small as 500 the total variation distance is typically below 0.05. For networks of size $n = 2500$ or more, the total variation distance is on average below 0.02. The rates for other values of k (not shown) are very similar to the case for $k = 3$.

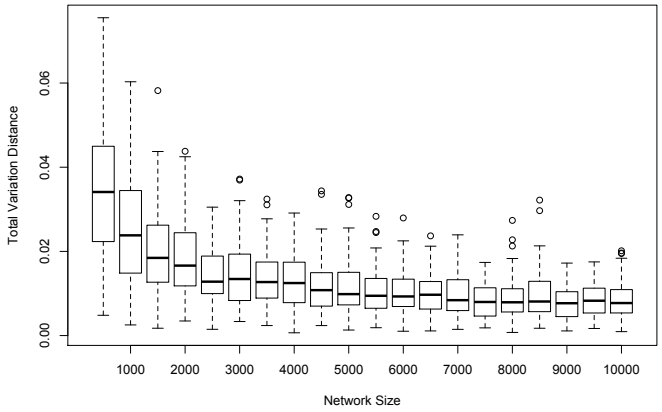

Figure 2: The total variation distance between a $Bin(3, .25)$ distribution and the number of edges from a vertex u to the collection B_n for all u with out-degree 3. For each size, the total variation distance is shown over 100 generations of a directed configuration model with in- and out- degree sequences which follow a power law distribution of exponent 3.

4.2 Analysis of Networks with Directed Community Structure

We now evaluate the effectiveness of our methodology on networks that contain prescribed directed community structure. We use the *stochastic co-blockmodel* of [19], a generative model that specifies the probability of connection between sender and receiver community (block) pairs. Given $r, p \in (0, 1)$ and $r+p \leqslant 1$, the stochastic co-blockmodel specifies that vertices of a sender community point (with outward directed edges) to vertices of an associated receiver community with probability $p+r$, while all other vertices point to one another with probability r. Note that one sender community can be associated with more than one receiver community and vice versa. Figure 3 shows an example of a stochastic co-blockmodel with one sender community A and two receiver communities B and C.

For our simulations, we generate stochastic co-blockmodels of size n with two equally sized communities A and B where vertices in one block point to vertices in the other with probability $p + r$, and vertices within the same block point to one another with probability r. We fix $r = 0.05$ and consider values of p corresponding to the signal to noise ratio $SNR = p/r$. We generate networks over SNR between 0 and 2 in increments of 0.025. At each SNR setting, we calculate the mean p-value of vertices in opposing blocks as well as the mean p-value of vertices within the same block. We repeat this simulation for networks of size $n = 100, 500, 1000$, and 2000. For each SNR value, we generate 30 networks and record the average p-value. We show the average p-value associated with the vertices of A in Figure 4. The distributions of p-values for the vertices from communities A and B to the collection B is shown for $SNR = 0.5$ in Figure 5.

We observe several important features of our p-value quantity from Figure 4. When $SNR = 0$, all vertices point to one another with equal probability r meaning that there is no preferential attachment between any vertex pair. Our p-values reflect that across all n, taking values around 0.5. As the value of the SNR grows, the strength of attachment of the vertices from A to B increases and the strength of attachment from the vertices in A to A decreases. This trend is captured by the trend of our p-values: the average $p(u : B)|_{u \in A}$ decreases to 0 and the average $p(u : A)|_{u \in A}$ increases to 1 as the SNR increases. The rate of convergence of these p-values increases as the size of the network increases. We further illustrate this point in Figure 5 by comparing the separation of $p(u : B)|_{u \in A}$

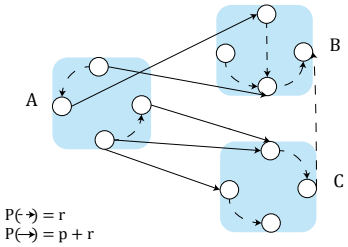

Figure 3: An example of the stochastic co-blockmodel with sender community A and receiver communities B and C. Here, edges point from A to both B and C with probability $p + r$ and all other vertices point to one another with probability r.

and $p(u : B)|_{u \in B}$ for various network sizes. For larger n, we see better separation of the two sets of p-values.

This example illustrates the effectiveness of our methodology on networks with prescribed community structure. In the case where all edges are randomly assigned with equal probability (when SNR = 0), the p-values hover around 0.5 correctly suggesting the network contains no significant local connections. As communities become more distiguishable (SNR $\geqslant 0.5$), the p-values appropriately capture the local structure. These observations suggest that these local p-values may be utilized as a community detection tool in its own right; however, we save this exploration for future work.

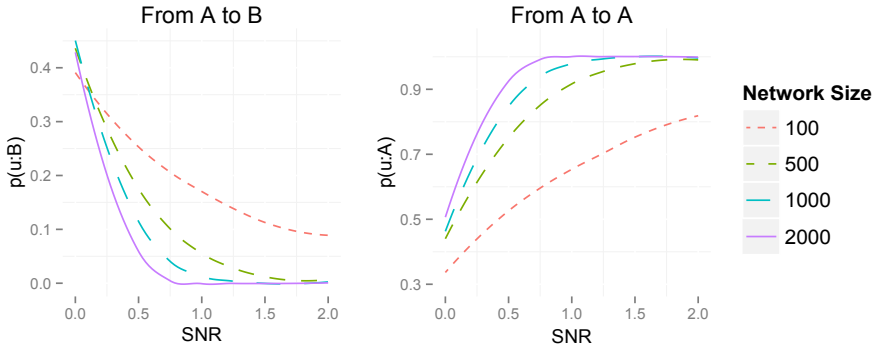

Figure 4: The strengths of connections of the sender community A to receiver communities A and B within the stochastic co-blockmodel simulations. The x-axis in each plot is the signal to noise ratio $SNR = p/r$ and the y-axis shows the average p-value of vertices from A to the given receiver community. We simulate the stochastic co-blockmodels as described in the text and illustrate these results for networks of size $n = 100, 500, 1000$, and 2000.

5 Political Blog Dataset

The political blog network of [1] is a snapshot of the webblog structure of 1494 political blogs on a single day closely following the 2004 U.S election. The vertices of the network represent the political blogs where each blog has been classified as either liberal or conservative by the authors in [1]. Directed edges represent hyperlinks from one blog to another in the network. We consider only blogs in the largest weakly connected component of this original network. The resulting network contains 1222 blogs - 636 liberal and 586 conservative - and 19021 directed edges.

We consider the two collections of vertices - the liberals and the conservatives - and calculate the strength of affiliation of every blog to these collections using our p-value scores. In words, the p-values quantify the extent to which each blog hyperlinks to the liberal and conservative groups. Unsurprisingly, we find that the blogs tend to link predominantly and significantly to blogs of their

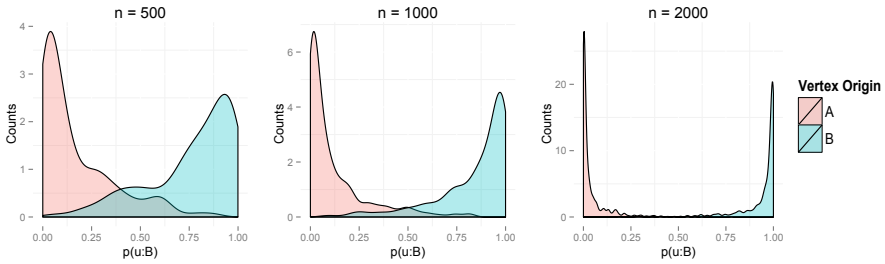

Figure 5: The distribution of p-values $p(u : B)$ for the generated co-block model at SNR = 0.5. From left to right, we show these distributions according to network size n. These figures suggest that as n grows, the communities A and B become more and more distinguishable.

own political affiliation, an observation which has been the primary result of numerous community detection studies [12, 18]. Our local methodology, however, provides additional insights about this hyperlink structure which we now describe. All quantitative results are shown in Table 1.

First of all, we observe that conservative blogs tend to make significant links, with p-value $\leqslant 0.10$, more often than their liberal counterparts. Moreover, the conservative blogs tend to more often link significantly to their own affiliated blogs. Indeed, 89% of the conservative blogs link significantly to at least one of the political groups in the network. Of these significant edges, 99% are made to their own conservative group. On the other hand, only 73% of the liberal blogs make significant connections from which 94.5% of the connections are to their own political group. These observations suggest that even though the liberal and conservative blogs tend to link predominantly to their own group, the linking tendency of the two groups are inherently different.

Next, our analysis identifies that a surprising 237 (19.4$\%$) of the blogs in this network do not make significant connections - as indicated by a p-value greater than 0.10 - to either the conservative or liberal groups. This suggests that these blogs may be one of two types: either one that shares roughly equal amounts of hyperlinks with both political groups, or one that is loosely connected and posts very few hyperlinks. Figure 6 illustrates this phenomenon in the network. Of these 237 blogs, 149 post fewer than 2 hyperlinks suggesting that these blogs are of the background type. Of the remaining 88 non-significantly connected blogs (32 conservative and 56 liberal), we find that the blogs still tend to have more connections with their own political affiliation. From Figure 7 we see that even when the blogs aren't strongly connected with either political party, the conservative blogs still tend to favor their own political affiliation more than their liberal counterpart. These results demonstrate that our p-values can shed light on overlapping and background community structure. These are two important features that arise in many real-world networks (see e.g, [23, 24, 26]) but are typically overlooked by community detection methods on directed networks.

Table 1: Political Affiliation of Webblogs (p-values < 0.10 suggest strong affiliation)

	Affiliated with		
	Conservative Group	Liberal Group	Neither
# Conservative Blogs (%)	514 (88)	8 (1)	64 (11)
# Liberal Blogs (%)	24 (4)	439 (69)	173 (27)
Total	538	438	237

6 Conclusion

We have proposed and investigated a methodology to assess the statistical significance of local connections from a vertex to any collection of vertices in directed networks. We have shown that under the directed configuration model, the number of directed edges from a vertex to a collection is approximately binomial. Through numerical simulations we have shown that this approximation is valid even for networks of only moderate size. Using this binomial distribution as a reference

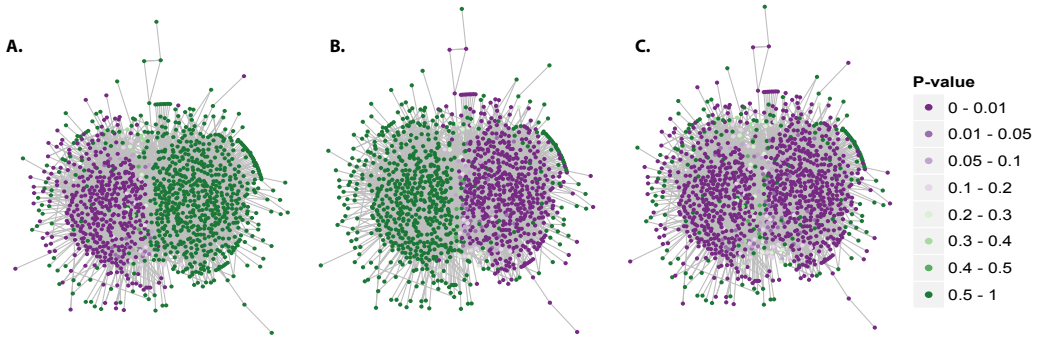

Figure 6: The p-values quantifying the strength of connection from each web blog to the groups of affiliated blogs. Plot **A** and **B** shows the p-values associated with edges from each blog to the collection of liberal and conservative blogs, respectively. Plot **C** shows the minimum p-value of connectedness to each collection of blogs. Plot **C** suggests that weakly connected blogs are either overlapping and close to the center of the two communities, or background and situated in the periphery of the network.

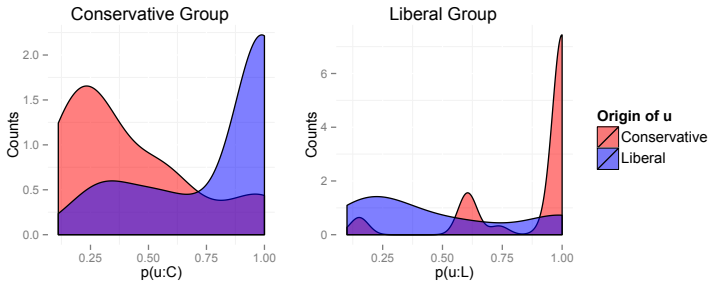

Figure 7: The p-values of the 88 weakly affiliated political blogs (minimum p-value > 0.10 and out-degree > 2). (Left): The p-value of association with the conservative group ($p(u : C)$). (Right): The p-value of association with the liberal group ($p(u : L)$). These figures suggest that even when the blogs aren't strongly affiliated with either group, they still tend to link to blogs of their own political affiliation.

measure, we quantify the significance of these directed connections through a p-value. We have shown when the network data are drawn from a model with prespecified community structure, these p-values readily identify community structure. Finally, we applied our methodology to a political webblog network where like community detection we find that the blogs tend to link to their own political group; however, only through analyzing the local structure of the blog network were we able to distinguish the linking tendencies of the two political groups and identify blogs that were not significantly linking to either group. Identifying local patterns in real-world directed networks and assessing their statistical significance is an important area of research that can be utilized in many real world applications. We have introduced one such methodology which provides an effective exploratory tool in this area.

Acknowledgments

This work was partially supported by NSF grant DMF-0645369, NSF grant DMF- 1105581, and NSF grant DMF-1310002.

References

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

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *Mixed membership stochastic blockmodels*, The Journal of Machine Learning Research **9** (2008), 1981–2014.

[3] A. Arenas, J. Duch, A. Fernández, and S. Gómez, *Size reduction of complex networks preserving modularity*, New Journal of Physics **9** (2007), no. 6, 176.

[4] E. A. Bender and E. R. Canfield, *The asymptotic number of labeled graphs with given degree sequences*, Journal of Combinatorial Theory, Series A **24** (1978), no. 3, 296–307.

[5] G. Bianconi, P. Pin, and M. Marsili, *Assessing the relevance of node features for network structure*, Proceedings of the National Academy of Sciences **106** (2009), no. 28, 11433–11438.

[6] B. Bollobás and A. Universitet, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, Aarhus Universitet, 1979.

[7] S. Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), no. 3, 75–174.

[8] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, *Social features of online networks: The strength of intermediary ties in online social media*, PloS one **7** (2012), no. 1, e29358.

[9] D. Greene, D. l. Doyle, and P. Cunningham, *Tracking the evolution of communities in dynamic social networks*, Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, IEEE, 2010, pp. 176–183.

[10] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, *Finding statistically significant communities in networks*, PloS one **6** (2011), no. 4, e18961.

[11] A. Lancichinetti, F. Radicchi, and J.J. Ramasco, *Statistical significance of communities in networks*, Physical Review E **81** (2010), no. 4, 046110.

[12] E. A. Leicht and M. E. J. Newman, *Community structure in directed networks*, Physical review letters **100** (2008), no. 11, 118703.

[13] Emmanuel D Levy and Jose B Pereira-Leal, *Evolution and dynamics of protein interactions and networks*, Current opinion in structural biology **18** (2008), no. 3, 349–357.

[14] F. D. Malliaros and M. Vazirgiannis, *Clustering and community detection in directed networks: A survey*, Physics Reports (2013).

[15] Julian McAuley and Jure Leskovec, *Learning to discover social circles in ego networks*, Advances in Neural Information Processing Systems 25, 2012, pp. 548–556.

[16] M. Molloy and B. Reed, *A critical point for random graphs with a given degree sequence*, Random structures & algorithms **6** (1995), no. 2-3, 161–180.

[17] M. E. J. Newman, *Detecting community structure in networks*, The European Physical Journal B-Condensed Matter and Complex Systems **38** (2004), no. 2, 321–330.

[18] M.E.J. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences **103** (2006), no. 23, 8577–8582.

[19] K. Rohe and B. Yu, *Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm*, arXiv preprint arXiv:1204.2296 (2012).

[20] V. A. Traag, G. Krings, and P. Van Dooren, *Significant scales in community structure*, arXiv preprint arXiv:1306.3398 (2013).

[21] A. L. Traud, P. J. Mucha, and M. A. Porter, *Social structure of facebook networks*, Physica A: Statistical Mechanics and its Applications **391** (2012), no. 16, 4165–4180.

[22] Y. J. Wang and G. Y. Wong, *Stochastic blockmodels for directed graphs*, Journal of the American Statistical Association **82** (1987), no. 397, 8–19.

[23] J. D. Wilson, S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel, *A testing based extraction algorithm for identifying significant communities in networks*, arXiv preprint arXiv:1308.0777 (2013).

[24] S. Xie, J.and Kelley and B.K. Szymanski, *Overlapping community detection in networks: the state of the art and comparative study*, arXiv preprint arXiv:1110.5813 (2011).

[25] J. Yang and J. Leskovec, *Defining and evaluating network communities based on ground-truth*, Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ACM, 2012, p. 3.

[26] Y. Zhao, E. Levina, and J. Zhu, *Community extraction for social networks*, Proceedings of the National Academy of Sciences **108** (2011), no. 18, 7321–7326.

[27] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf, *Learning from labeled and unlabeled data on a directed graph*, Proceedings of the 22nd international conference on Machine learning, ACM, 2005, pp. 1036–1043.

Supplemental Material

Approximate Distribution of $\hat{d}(u:B)$

Here we prove Theorem 1 which gives the approximate law of $\hat{d}(u:B)$ from which our proposed p-values are derived. This result is specific to the directed configuration model which is used as a null network network model for quantifying statistically significant connections.

Proof: Equation (3) implies that for the total number of edges $|E_n|$ one has

$$\int_{\mathbb{R}} x dF_n(x) = \sum_{k=0}^{\infty} k \frac{N_k(n)}{n} = \frac{|E_n|}{n} \sim \mu$$

where $N_k(n)$ is the number of vertices of in-degree k. Thus $|E_n| \sim n\mu$.

Consider the distribution of $\hat{d}(u:B)$, namely the number of connections from vertex u to the subset B in DCM($\{\mathbf{d}_o^{(in)}, \mathbf{d}_o^{(out)}\}$) on a vertex set $[n]$. We drop the n notation throughout for simplicity. When constructing the directed configuration model, one can start at any vertex and start sequentially attaching its outward half-edges at random to available inward pointing half edges. Thus we start with the fixed vertex u and decide the inward pointing half edges paired to the $d^{out}(u) := k^{out}$ outward half edges of vertex u. Let A_1 be the event that the first outward pointing half-edge of vertex u connects to the collection of vertices B. Let $p_1(B)$ denote the probability of this event. Then,

$$p_1(B) = \frac{\sum_{v\in B} d^{in}(v)}{[\sum_{v\in[n]} d^{in}(v)] - 1} = \frac{\sum_{v\in B} d^{in}(v)}{|E| - 1} \tag{7}$$

In general for $1 \leqslant i \leqslant k^{out}$, let A_i denote the event that an outward half-edge i of u connects to the set B and write $p_i(B)$ for the conditional probability of A_i conditional on the outcomes of the first $i-1$ choices. For $i = 2$, we claim that uniformly on all outcomes for the first edge, this conditional probability can be bounded as

$$\frac{[\sum_{v\in B} d^{in}(v)] - 1}{|E| - 2} \leqslant p_2(B) \leqslant \frac{\sum_{v\in B} d^{in}(v)}{|E| - 2} \tag{8}$$

The lower bound in (8) arises if the first outward half-edge of u connected to a half-edge pointing to B while the upper bound arises if the first outward half-edge does not connect to a half-edge pointing to B. Arguing analogously for $1 \leqslant i \leqslant k$ we find that the conditional probability $p_i(B)$ that the i-th half-edge of vertex u connects to B is bounded (uniformly on all choices of the first $i-1$ edges) as

$$\frac{[\sum_{v\in B} d^{in}(v)] - (i-1)}{|E| - i} \leqslant p_i(B) \leqslant \frac{\sum_{v\in B} d^{in}(v)}{|E| - i} \tag{9}$$

Recall that $p(B) = \sum_{v\in B} d^{in}(v)/|E|$. Using (9) and the fact that $|E| \sim n\mu$, we have

$$\sup_{1\leqslant i\leqslant k^{out}} |p_i(B) - p(B)| \leqslant \frac{k^{out} - 1}{|E| - k^{out}} \to 0 \tag{10}$$

as $n \to \infty$.

Finally, note that the random variable $\hat{d}(u:B)$ can be expressed as

$$\hat{d}(u:B) = \sum_{i=1}^{k^{out}} \mathbb{1}\{A_i\}$$

Thus, using (10) we have that

$$d_{TV}(\hat{d}(u:B), \text{Bin}(k, p(B))) \to 0, \qquad \text{as } n \to \infty.$$

This completes the proof.

■