Comparing predictive powers of Network Motif Distribution and structure of Overlapping Communities

Ling-Ling Zhang, Will Thomas, Awo Ashiabar

Abstract

Arguably, network motif distribution and the overlapping structure of a network can inform on the topology of a network and, more importantly, on the function of a network. In this paper, we explore if network motif distribution and overlapping structures can explain and distinguish between two functionally dissimilar social networks: Facebook, an online representation of social communities, and Twitter, a platform for celebrity-like figures to broadcast to many followers. We apply exhaustive and stochastic algorithms to collect counts of 3-node and 4-node open and closed motifs in both networks. In addition, we test three overlapping communities detection algorithms, Mixed Membership Stochastic Blockmodel, AGM and BIGCLAM, on the networks. In order to understand observations that are statistically significant, the results of the social networks are compared to that of randomized networks. Comparing results of the test networks against randomized networks elucidates the motif signatures and structure of overlapping communities unique to the two real social networks.

Related Work

Motifs:

In the paper, *"Network Motifs: Simple Building Blocks of Complex Networks"*, R. Milo, et al. map various 3-node and 4-node motifs in complex networks and discover that certain motifs occur more frequently in particular types of graphs. For example, Milo implies that certain feed-forward 3-node motifs are over-abundant in information processing networks and are likely to represent information processing functions. Similar to the work of Milo et al, we map out several 3-node and 4-node motifs in the two social networks to understand if the differences in motif distribution between the Twitter and Facebook networks correspond to the broadcast function of Twitter and the close-knit personal relations on Facebook.

Our project closely resembles work completed by Turkett et al [2] in "*Graph mining of motif profiles for computer network activity inference"* where Turkett's team by comparing the motif distribution of computer traffic networks identify with 85% accuracy seven types of applications including AOL Instant Messenger (AIM), Hypertext Transfer Protocol (HTTP), Domain Name System (DNS), Kazaa, Microsoft Active Directory Domain Services (MSDS), NetBIOS Name Service, and Secure Shell (SSH). The test run by Turkett et al [2] is compromised in that the researchers do not know for certain the applications under tests. This project avoids this shortfall because the applications of the networks under tests are known apriori and measurements of the predictive power of motif distributions and overlapping communities are indisputable.

Overlapping Community Structure:

"*Uncovering the Overlapping Community Structure of Complex Networks in Nature"* by Palla et al. explores the notion that networks in complex systems are are more than just dense, well-separated

communities, but rather groups of highly overlapping nodes. Rather than characterizing nodes as belonging to any single cluster (as one might when considering the clustering coefficient of a graph), Palla et al. characterizes each node as belonging to multiple communities, and communities not as being separate but as being possible overlapping and/or nested. An advantage of this model is that it does not force a node to belong to a specific category, thus preventing communities from being broken up during the process of classifying nodes.

In "*Mixed Membership Stochastic Blockmodels*" by Airoldi et al, the notion of using a mixed membership model to approach relational data in order to find overlapping community structure. The mixed membership model associates each unit (a person, a word, etc.) with multiple clusters (social circles, themes, etc.). They showed through various examples how this model was able to accurately capture different social groups as well as the "wavering" or "uncertain" affiliations of people with respect to those social groups. In this work we test the mixed membership model by running the R implementation of the algorithm.

In "Community-Affiliation Graph Model for Overlapping Network Community Detection" by Yang et al, they explore the structure of overlapping networks. They observed that the areas of the graph where communities overlap are more densely connected than non-overlapping parts, and created a model AGM, or Community-Affiliation Graph Model to improve on overlapping community detection.

In "Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach" by Yang et al, they improve on large-scale community detection using matrix factorization. Using a bipartite affiliation network model they learn factors that allow them to recover the underlying adjacency matrix of the network. Unlike traditional uses of Nonnegative Matrix Factorization this paper aims to learn the community affiliation of nodes that explains the links in the observed network.

Model/Algorithm/Results

Motifs:

To compute the motif distributions, we use the *mfinder* tool (by Milo et. al.) for exact motif counts and a stochastic algorithm (by Kashtan et. al.) for accurate but quick motif assessments. The *mfinder* tool uses a brute force approach to count the exact concentrations of different motif types. For each node, it considers all edges extending from the starting node. It follows each of these edges in turn to generate a community inclusive of the starting node. By iterating over all the nodes, *mfinder* counts all motifs in the graph. However, due to the size of the Facebook and Twitter networks especially the Twitter network which has an excess of 1M edges, we resorted to the sampling algorithm by Kashtan et al . The stochastic algorithm works by repeatedly sampling random edges in a graph and expanding the community sampled by adding edges incident to the current subgraph. It continues populating the community in this way until it generates a motif of a desired size. This sampling procedure does not detect all motifs with uniform probability, consequently, the resulting sampling count is weighted. Motifs that are more likely to be sampled are weighted less. Run time of the sampling algorithm depends only on number of samples and is independent of the size of the network.

Overlapping Community Structure:

Mathematical Background

The Mixed Membership Stochastic Model makes heavy use of various distributions including the Dirichlet, Multinomial, and Bernoulli distributions.

The Dirichlet distribution is a n-dimensional vector which is the conjugate prior of the parameters of the multinomial distribution (the multinomial distribution is explained below). This means that given a model where the prior distribution of a data point is Dirichlet, and the data point has a multinomial distribution, the posterior distribution is also Dirichlet.

The multinomial distribution gives the probability of choosing a set of m specific items from a set of n items when picking with replacement.

The Bernoulli distribution is a discrete probability distribution with two cases, “success” and “failure” where p usually denotes the probability of success, and q = 1 - p the probability of failure.

BIGCLAMP makes use of NMF or Nonnegative Matrix Factorization. NMF takes in a matrix V and attempts to construct two matrices which multiply to V. What makes this problem slightly more tractable is the fact that all three matrices must consist of only non-negative values.

MMSB

We explore three different algorithms to detect overlapping community. The first is the MMSB algorithm. The second is the AGM, or Communitiy-Affiliation Graph Model. The third is BIGCLAM, or Cluster Affiliation Model for Big Networks.

The MMSB algorithm is given two parameters N and k where N is the number of units and k is the number of groups, and a binary matrix Y where Y_{ij} indicates positive interaction (for instance i follows j on twitter). α and B are fixed within the algorithm. The algorithm then runs as follows:

π_i = Dirichlet(α)
$z_{i \to j}$ = Multinomial(π_i)
$z_{j \to i}$ = Multinomial(π_j)
for i in range N
    for j in range N
        A_{ij} = $z_{j \to i}$ π_i $z_{i \to j}$
Y' = Bernoulli (A_{ij}, Y)
MCMC(π_i, $z_{i \to j}$, $z_{j \to i}$, Y', α, B)

Where MCMC is the Markov Chain Monte Carlo algorithm.

So far we have modified the algorithm to plot the π_i, $z_{i \to j}$, and $z_{j \to i}$ vectors along with summary statistics for π_i, $z_{i \to j}$, and $z_{j \to i}$ and Y. In the future we plan to implement an approximation algorithm for choosing the number of groups given a graph. This algorithm will be based off the BIC approximation presented in “*Mixed Membership Stochastic Blockmodels*” by Airoldi et al.

AGM

AGM is a probabilistic generative model for graphs that captures the overlapping structure of the network using community affiliations. The main ideas behind this model are as follows:

- communities arise from shared group affiliations
- people tend to be involved in communities to various degrees

This gives rise to a bipartite affiliation network as it's model for communities in a graph. The nodes on one side represent the nodes in some network G, while the nodes on the other side represent communities. An edge on the graph indicates membership. Given a bipartite graph B, for each community c we assign a parameter p_c, the probability of an edge forming between two members of c. We then have each community generate edges independently between it's members.

Thus to detect communities, AGM attempts to fit an unlabeled undirected graph G(V,E) to the AGM model by finding the graph B and parameters p_c that maximizes the likelihood that they generate G.

BIGCLAM

BIGCLAM builds off much of the same ideas underlying AGM with one additional observation:

- when people share multiple affiliations, they are more likely to belong to the same community

Like the AGM model, the BIGCLAM model uses a bipartite affiliation network to model the underlying community structure in the graph. Given a bipartite graph B we generate a nonnegative matrix F . Then for every pair of nodes u, v, we connect them with probability $p(u,v)$ where

$$p(u,v) = 1 - \exp(-F_u \cdot F_v{}^T)$$

To detect communities BIGCLAM, like AGM, attempts to fit an unlabeled undirected graph G(V,E) to the BIGCLAM model with K communities by maximizing the likelihood that F generates G or:

$$\max l(F),\ l(F) = log\, P(G|F)\,.$$

But this still requires us to specify a value of K which is undesirable. Thus in order to discover what value of K is, BIGCLAM runs it's initial pass on 80% of the graph and uses the remaining 20% as validation. It does this for several values of K and picks the value that maximizes likelihood.

Results and Findings

Motifs:

For brevity, we refer to the different 3-node and 4-node motifs by id values. See Figure 1 for a visual representation of the relevant motifs.

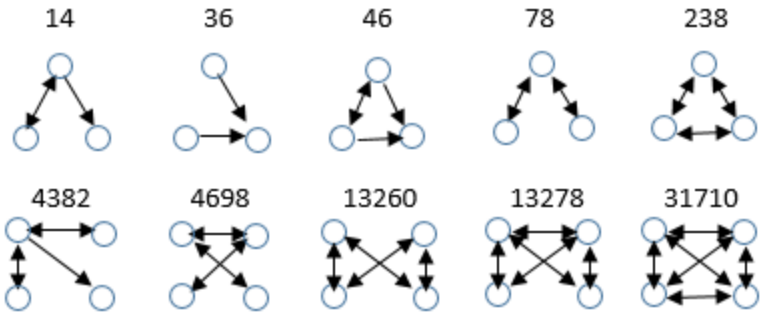

Figure 1. 3-node and 4-node subgraphs and their motif ids

Motif Distribution :

In this experiment, we first test the stochastic algorithm by comparing exact 3-node motif distributions of the Facebook and Twitter networks to stochastic measurements using 10,000 samples. The difference between the exact and stochastic results are almost identical for high volume motifs. See Table 1.

Motif ID	Actual Concentration	Sampling Concentration	Percent Error
14	0.260	0.361	38.85
46	0.610	0.955	56.56
36	18.700	16.221	13.26
238	261.710	263.544	0.701

Table 1 shows exact and stochastic motif measurements. Differences are negligible for motifs in high concentrations. Concentration values are scaled by 1,000 for legibility.

Before settling on using 10,000 samples for the remainder of the exercise, we test the accuracy of the stochastic algorithm for various numbers of samples and discover that 10,000 samples provided accurate readings for highly concentrated motifs. The error for the low concentrated motifs (less than 100), is relatively high but excusable since the goal of the study is to analyze highly concentrated motifs. Indeed, the error of the two most abundant motifs is less than 1%.

Motif ID	Number of Samples				
	5,000	10,000	15,000	20,000	25,000
14	0	0.361	0	0.019	0.014
46	0.702	0.955	0.598	0.517	0.861
36	16.381	16.221	22.230	22.442	22.455
238	267.392	263.544	265.068	260.777	257.631
78	715.525	718.919	712.005	716.245	719.038

Table 2. Distribution of 3-node motifs in Facebook networks by various sample sizes.Concentration values are scaled by 1,000 for legibility.

Comparing Motif Distributions of Facebook, Twitter and Random Networks:
Table 3 compares the motif distributions of the Facebook and Twitter networks to random networks. The appendix Tables Appx-1 to Appx-4 also display distributions for other randomized graphs fit as Kronecker graphs and Erdos-Renyi.

Motif Type	Motif ID	Facebook (FB)		Twitter (TW)	
		Real Network	Randomized with FB Degree Distribution	Real Network	Randomized with TW Degree Distribution
3-node	14	0.361	6.576	0.662	1.161
	46	0.955	2.074	0.775	51.66
	36	16.221	125.747	0	0
	238	263.544	25.661	936.901	933.802
	78	718.919	839.942	61.663	13.377
4-node	2462	0.059	1.036	694.013	471.806
	3038	0.01	0.009	220.117	338.531
	13260	8.555	6.97	2.108	1.694
	13278	70.715	3.866	7.179	2.164
	31710	44.58	0.177	1.339	0.293

Table 3. Distribution of 3-node and 4-node motifs in Facebook and Twitter networks. Concentration values are scaled by 1,000 for legibility.

Both Facebook and Twitter networks exhibit a much higher concentration of the motifs ids 238 (complete 3-node sub-graph) and 31710 (complete 4-node sub-graph) compared to random graphs. Even the randomized graphs with similar degree distribution have fewer complete sub-graphs.

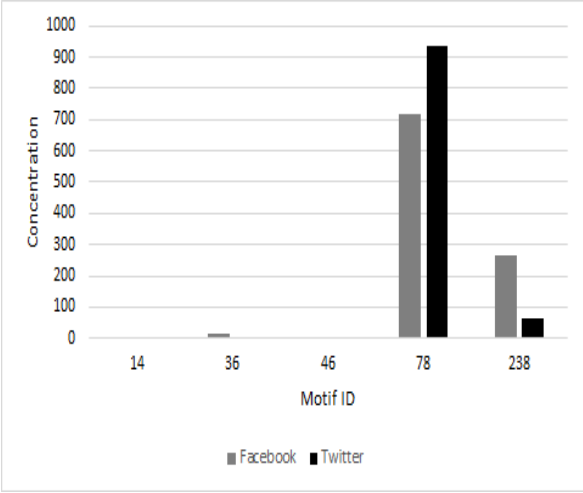

Figure 2. Size-3 Motifs in Facebook and Twitter networks

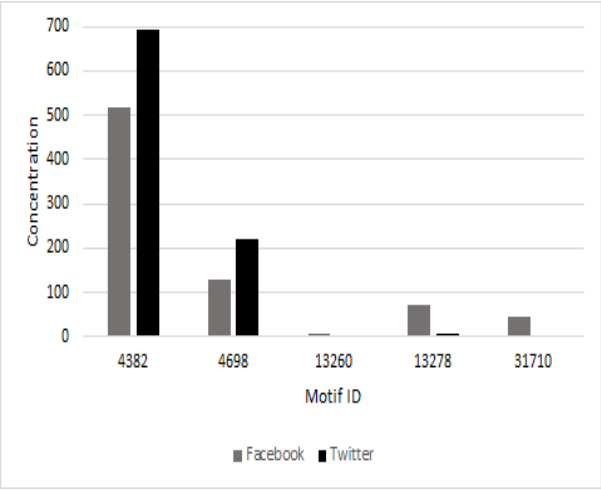

Figure 3. Size-4 Motifs in Facebook and Twitter networks

Specifically on the subject of differences between Facebook and Twitter, the concentrations of the complete motifs are more prevalent in the Facebook network than in the Twitter network. The concentration of motif 238 (3-node complete sub-graph) in the real Facebook network is 4 times higher than that of the real Twitter network. Similarly, the concentration of motif 31710 (4-node complete graph) is 33 times more abundant in the real Facebook network than in the real Twitter network. These results are consistent with expectations. As a mirror of real human relations, we expect communities on Facebook to be dense and cross-connected. Conversely, communities on Twitter network primarily serve a broadcast function and although we expect some connections to be interwoven, Twitter is a human network afterall, Twitter connections are expected to be relatively sparse. Users are more likely to broadcast to subscribers outside their immediate personal relationships, i,e, users are more likely to broadcast to users who share few connections with other users following the same broadcast. It is worth noting, however, that the concentrations of the complete motifs are still higher in the Twitter network than in random networks.

Overlapping Community Structure:

Mixed Membership Model

The two types of networks were categorized under 5 communities. In the table below we show the distribution of overlapping communities of both networks.

Distribution of overlapping networks of Facebook and Twitter networks.				
Number of Overlapping Communities	**Facebook**		**Twitter**	
	Real	**Randomized with same Degree distribution**	**Real**	**Randomized with same Degree distribution**
1	0.9%	1.4%	0.0%	0.1%
2	3.8%	3.9%	0.5%	0.7%
3	19.0%	20.9%	11.4%	11.9%
4	53.9%	53.6%	50.9%	50.9%
5	22.3%	20.3%	37,1%	36.4%

Table 4. A node is said to belong to a community if π , the proportion of times a node substantiates its relationship with a group, >0.15

Compared to the Twitter network, only 22% versus 37% of the Facebook network belong to all 5 identified communities. This result makes intuitive sense. On Twitter, an umpteen number of people follow a handful of celebrities and if the 5 categorized communities on the test Twitter network are analogous to the following of the top 5 celebrities, chances are a good percentage of the Twitter network follow all top 5 celebrities given the many-to-one connections on Twitter. Facebook on the other hand exemplifies the “Small World” concept where friends form closed connections to friends of friends. Naturally, Facebook has many closed triangles and by extrapolation many closed disjoint sub-communities. It is not surprising therefore that on Facebook relatively fewer nodes belong to all 5 communities. What is surprising is that the overlapping community structure of the real social networks and randomized graphs are identical. This implies that unlike network motif distribution, structures of overlapping communities cannot enable the detection of differences in social network that bear similar degree distributions.

In the table below we compare the AGM and BIGCLAM model on several facebook ego-networks. We could not run it on the entire facebook network or any twitter ego-networks due to the limited scale of graphs AGM could run on. We did not include results from the original MMSB algorithm as it requires the number of communities as input.

	Ground Truth	AGM	BIGCLAM
Facebook0	24	6	22
Facebook414	6	2	6
Facebook686	13	7	16
Facebook1684	17	8	20

What we discovered is that BIGCLAM is a much more accurate model than AGM is on detecting communities in networks. AGM consistently underestimated the number of communities, often collapsing several communities into one, whereas BIGCLAM as able to retain much of the original communities. As a result, we chose to use BIGCLAM to do our comparison of the Facebook and Twitter ego-networks as well as the random graphs we generated.

In the table below we compare statistics obtained from BIGCLAM on the Facebook and Twitter ego-networks as well as the random graphs we had generated.

	Avg. group size	Avg. membership	Avg. membership w/out ungrouped nodes
Facebook	28.35	.98	1.34
Twitter	13.7	1.01	1.43
Random Facebook	527.8	1.58	1.58
Random Twitter	3425.4	1.55	1.55

Here we see that Twitter networks have a significantly smaller group size than Facebook networks. Twitter networks are representative of personal interests rather than group affiliation, incentivizing increased diversity between people.. In a celebrity network like Twitter people have less incentive to follow every person they know in real life. Rather, they only follow people they find interesting or funny, (whereas facebook goes out of it's way to recommend friends to people based on mutual friendships). As a result there are more differences between individual twitter nodes than individual facebook nodes which are more constrained due to location.

We also found that randomly generated networks, even with the same degree distribution as the original facebook and twitter networks, do not exhibit community structures that would be present in real networks. This is relatively unsurprising since there this model does not favor the creation of "communities" (creating

more edges within a group of nodes) but instead does a random global creation which is conducive to creating big global communities, but not several smaller overlapping or nested ones.

Conclusion

The results of this study support Milo's suggestion in *"Network motifs: simple building blocks of complex networks"* that network motifs inform on the function of networks. In the Facebook network where persons are interconnected in close communities, we find an over-abundance of complete subgraphs than we do in random graphs or Twitter network.

Although the initial results in differentiating Facebook and Twitter using Overlapping Networks is promising, more samples would be needed to see if they can differentiate between friendship and celebrity networks in general. The discrepancy in the average size of Facebook and Twitter communities, while significant, is not necessarily due to the differences between friendship and celebrity networks. On the other hand, there does appear to be a very clear difference in community structure between real-world networks, and networks that were generated with the same degree distribution.

Appendix

Distribution of 3-node motifs of Facebook networks				
Motif ID	**Real Network**	**Randomized with Same Degree Distribution**	**Erdos-Renyi Graph**	**Kronecker Graph**
14	0.361	6.576	1.046	5.315
46	0.955	2.074	0	0.257
36	16.221	125.747	0.838	15.140
238	263.544	25.661	3.441	11.07
78	718.919	839.942	994.675	968.217

Table Appx-1. 3-node motif distribution of Facebook real and random graphs. All concentration values are multiplied by 1,000 for convenience.

Distribution of 4-node motifs of Facebook networks				
Motif ID	**Real Network**	**Randomized with Same Degree Distribution**	**Erdos-Renyi Random Graph**	**Kronecker Graph**
2462	0.059	1.036	0	0.098
3038	0.01	0.009	0	0
13260	8.555	6.97	2.077	4.141

13278	70.715	3.866	0.076	1.038
31710	44.58	0.177	0	0.006

Table Appx-2. 4-node motif distribution of Facebook real and random graphs. All concentration values are multiplied by 1,000 for convenience.

Distribution of 3-node motifs of Twitter networks				
Motif ID	**Real Network**	**Randomized with Same Degree Distribution**	**Erdos-Renyi Random Graph**	**Kronecker Graph**
14	0.662	1.161	0	0
36	0.775	51.66	0	0
46	0	0	0	0
78	936.901	933.802	999.858	998.449
238	61.663	13.377	0.142	1.551

Table Appx-3. 3-node motif distribution of Twitter real and random graphs. All concentration values are multiplied by 1,000 for convenience.

Distribution of 4-node motifs of Twitter networks				
Motif ID	**Real Network**	**Randomized with Same Degree Distribution**	**Erdos-Renyi Random Graph**	**Kronecker Graph**
4382	694.013	471.806	248.226	533.029
4698	220.117	338.531	750.306	458.553
13260	2.108	1.694	0.07	0.338
13278	7.179	2.164	0	0.098
31710	1.339	0.293	0	0

Table Appx-4. 4-node motif distribution of Facebook real and random graphs. All concentration values are multiplied by 1,000 for convenience.

References

[1]R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan,D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks.Science, 298(5594):824–827, 2002

[2]W. Turkett Jr, E. Fulp, C. Lever, and E. Allan Jr, “Graph mining of motif profiles for computer network activity inference,” in Ninth Workshop on Mining and Learning with Graphs, 2011.
[3]G. Palla, I. Derenyi, I. Farkas, T. Vicsek. “Uncovering the Overlapping Community Structure of complex networks in nature and society”, Nature 435, 814-818, 2005.
[4]E. Airoldi, D. Blei, S. Fienberg, E. Xing. “Mixed Membership Stochastic Blockmodels,” Journal of Machine Learning Research 9 (2008) 1981-2014.
[5]J. Yang, J. Leskovec. “Community-Affiliation Graph Model for Overlapping Network Community Detection”, ICDM ’12, 2012.
[6]J. Yang, J. Leskovec. “Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach”, ACM International Conference on Web Search and Data Mining (WSDM), 2013.

Contributions

Will: calculation of motif distributions and analysis of the motif data for the report
Ling-Ling: problem formulation, overlapping network comparison formulation, research/ write-up of of MMSB, AGM and BIGCLAM algorithms. Execution, analysis, and result write-up of AGM and BIGCLAM algorithms.
Awo: random graph creation, execution of MMSB algorithm (including input modification) and analysis of MMSB results, report write-up