Temporal Motif Degree Vectors
Efficient Mesoscale Characterization of Temporal Graphs

Benjamin Hannel, bhannel@stanford.edu

December 8, 2018

1 Introduction

Graph theory provides a common set of tools to analyze social networks, computer networks, protein interaction networks, financial networks, as well as many other types of real world processes in which entities relate to each other. Many real world networks also contain additional information characterizing their nodes and edges. In particular, in temporal networks, each edge is labeled with a real number, representing the time the two nodes interacted.

One way to understand the structure of networks is to look at the prevalence of small (3 to 8 node) motifs and their rate of occurrence within the network. This can provide insight about the structure of the network as a whole (e.g. it contains an unexpectedly large number of triangles), or about local structure in particular subgraphs. Earlier work has generalized this technique of characterizing graphs with motif frequencies to temporal graphs[3].

1.1 Present work

Earlier work has primarily focused characterizing entire graphs with temporal motifs. However, counting the motifs within the entire graph does not allow one to differentiate the role and local structure of individual nodes within the graph. To this end, we extend the graphlet degree signature technique[5] to temporal motifs in an effort to characterize the local structure of each node. We can use this information to demonstrate how the local structure around subpopulations of nodes differs from typical nodes. On dense graphs, computing this degree vector can be expensive, so we introduce a sampling technique to approximate it and prove a bound on the estimator.

2 Related Work

We combine and expand upon earlier work. We extend Milekovic et. al's notion of a graphlet degree vector to temporal graphs, though we do not require graphlets be induced. We examine two different definitions of temporal motifs, and opt to use the one more appropriate to the financial network context.

2.1 Uncovering Biological Network Function via Graphlet Degree Signatures

Milenkovic uses the general technique of motif counting to characterize the local neighborhood of a single node within a larger static (non-temporal) network [5]. Instead of counting motifs (or graphlets) over the entire network, the authors count only the number of instances of motifs which contain the node of study (ego). They also record where in the motif ego occurs, accounting for symmetry to remove redundancy. They successfully use the technique to identify distinct patterns in a food web, and discover protein complexes within the protein-protein interaction network.

2.2 Temporal Motifs

Kovanen et. al introduce the framework of a temporal motif[3]. A temporal motif in their paper is a set of k events (or edges) over a subgraph of n nodes, where every pair of edges is Δt-connected. The demonstrate an algorithm to efficiently count temporal motifs by this definition. In a graph of cell phone calls, the most common motifs are between two nodes (i.e. A calls B and B calls back, etc) and most motifs resemble causal chains.

2.3 Motifs in Temporal Networks

Paranjape, Benson, and Leskovec describe another fast algorithm for counting temporal motifs, using a different definition than Kovenan et. al [1]. Paranjape et. al's definition does not require that all of the events for a given node in a motif be consecutive for that node, which they postulate better captures important network events and structure. Indeed, this definition seems better suited to a financial network. Consider for example a firm which takes payment A for a order, then makes payment B to acquire materials to fulfill that order. The firm may make or receive other payments $C, D, ...$ in the interval of time between B and C, but B and C are still causally connected. Therefore it makes sense to count financial motifs based on a time window constraint, not a consecutive edge constraint. However, the algorithm presented by Paranjape et. al does not generalize very well to motifs larger than 3 nodes, and it handles many special cases separately in a complicated way.

2.4 Analytical Null Models for Temporal Motifs

In order to determine the significance of motif counts in any graph, one must compare it to the distribution of motifs in some null model. If the distribution in the actual graph is significantly different in some way from the null model, it can be claimed that the null model does not describe the true generative structure of the graph accurately. However, picking an overly simplistic null model, like Erdős-Renyi or configuration models, can often lead to exaggerated significance values which ultimately indicate nothing; few rare world processes are expected to resemble these simplistic models. An appropriate model can be used to test a more specific hypothesis. Mirzasoleiman in her paper defines several null models for temporal networks; constant edge arrival rates, dynamic edge arrival rates, and a stochastic block model [2]. She also calculates analytically the expected motif distributions of these models so that real graphs can be compared against them without computationally costly simulation. She validates these techniques against real world data sets, including the financial transaction network studied in this paper. For example, the motif distribution of the financial network, as compared to the stochastic block model as a null hypothesis, very clearly changes during the September 2011 financial crash.

3 Preliminaries

Definition 3.1. A *temporal graph* consists of a set of nodes V and a set of edges E where each edge in $e_i \in E$ is a 3-tuple consisting of the source node, the destination node, and the timestamp of the edge. Let all timestamps t_i be unique.

$$e_i = (u_i, v_i, t_i), u_i \in V, v_i \in V, t_i \in \mathbb{R}$$

These edges together form a directed multigraph where each edge is labeled with a real number.

Definition 3.2. A *k-edge temporal motif* is an ordered sequence of k edges.

$$(u_1, v_1, t_1), (u_2, v_2, t_2), ..., (u_l, v_l, t_k), t_1 < t_2 < ... < t_k$$

The static subgraph graph containing the edges $(u_1, v_1), ..., (u_k, v_k)$ and the nodes $\{u_1, v_1\} \bigcup ... \bigcup \{u_k, v_k\}$ must also be weakly connected.

Definition 3.3. A *δ-instance of a k-edge temporal motif* is a k-edge temporal motif for which all of edges are contained within a window of δ time.

$$t_1 + \delta \geq t_k$$

Definition 3.4. Automorphism orbit of a temporal motif: Let M and M' be k-edge template temporal motifs.

$$M = (u_1, v_1, t_1), (u_2, v_2, t_2), ..., (u_l, v_l, t_l), t_1 < t_2 < ... < t_k$$

$$M' = (u'_1, v'_1, t'_1), (u'_2, v'_2, t'_2), ..., (u'_l, v'_l, t'_l), t'_1 < t'_2 < ... < t'_k$$

M' and M are isomorphic if there exists a bijection between the nodes of M and M' f such that $f(u_1) = u'_1, f(v_1) = v'_1, ...$

A pair of nodes $n \in M$ and $n' \in M'$ occupy the same *automorphism orbit* if $f(n) = n'$. For example, u_1 and u'_1 occupy the same automorphism orbit. This equality operator is transitive between nodes in different motifs, so from now on we will refer to a node as being an instance of an automorphism orbit without comparing it to any other nodes.

Note that temporal motifs, unlike static motifs, are never automorphic, so each node in a motif occupies a unique automorphism orbit.

Definition 3.5. The *temporal motif degree vector* of a node is constructed by enumerating all δ-instance k-edge temporal motifs the node appears in, and counting how many times the node appears in each automorphism orbit of each motif. These counts are concatenated together into a vector with a consistent ordering.

4 Method

Here we propose an algorithm for computing the temporal motif degree vector of a given node. Using these vectors, we can compare the vector for a given node or subset of nodes to the distribution of vectors for all nodes. We can also divide the graph into time slices to see how the motif distribution changes over time. These approaches have the virtue that they do not rely on a null model. Earlier work has found the use of a null model for motif counts problematic [6] because null models must be very carefully designed to test a given hypothesis about how a network is structured. Here we compare subpopulations of nodes in the graph to other subpopulations, eliminating the need for a finicky null model.

4.1 Temporal Subgraph Enumeration

The algorithm is inspired by the exact subgraph enumeration algorithm [4]. However, as temporal graphs are in general multigraphs, we adapted it to recursively build up the edge set, rather than the vertex set. Two edges are considered adjacent if they share a common vertex.

```
function ExtendMotif($G$, $k$, $\delta$, $E_{subgraph}$, $E_{ext}$, $E_{adjacent}$)
    if IsDeltaInstance($E_{subgraph}$, $\delta$) then
        return
    end if
    if $|E_{subgraph}| = k$ then
        ProcessMotif($E_{subgraph}$)
    end if
    while $|E_{ext}| > 0$ do
        $e = $ Pop($E_{ext}$)
        $u, v, t = e$
        $E'_{ext} = $ (EdgesOf($G$, $v$) $\bigcup$ EdgesOf($G$, $u$) $/E_{adjacent}$)
        ExtendMotif($G$, $k$, $\delta$, $E_{subgraph} \bigcup e$, $E_{ext} \bigcup E'_{ext}$, $E_{adjacent} \bigcup E'_{ext}$)
    end while
end function

function EnumerateTemporalMotifs($G$, $k$, $v$, $\delta$)
    $E_{ext} = $ EdgesOf($G$, $v$)
    $E_{adjacent} = $ Copy($E_{ext}$)
    ExtendMotif($G$, $k$, $\delta$, $\{\}$, $E_{ext}$, $E_{adjacent}$)
end function
```

- G is the graph in which we are counting motifs
- k is the number of edges in each motif
- v is the target node. That is, every found motif must contain this node.
- δ is the time window the motif must fall in
- $E_{subgraph}$ is the set of edges added to the motif thus far
- E_{ext} is the set of edges which are adjacent to an edge in $E_{subgraph}$ and eligible to be the next edge added
- $E_{adjacent}$ is the set of all edges adjacent to $E_{subgraph}$, including $E_{subgraph}$

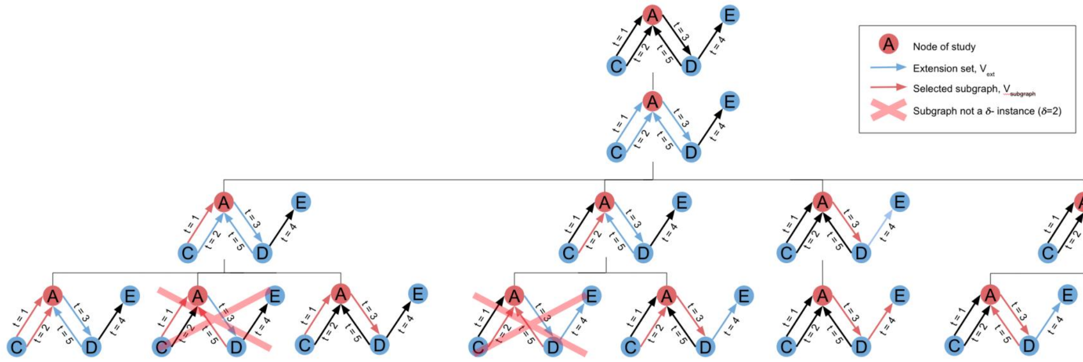
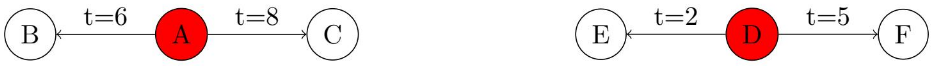

A
C
D
E
t=1
t=2
t=3
t=4
t=5
Node of study
Extension set, V_ext
Selected subgraph, V_subgraph
Subgraph not a δ- instance (δ=2)

Figure 1: An example execution of temporal subgraph enumeration for $k = 2, \delta = 2$

4.2 Temporal Subgraph Isomorphism

Graph isomorphism is in general a hard problem. However, on temporal graphs there is an additional constraint that for two graphs to be isomorphic, the edges must occur in the same order. This creates a readily available bijection between the edges of any pair of graphs, and because the graph is directed, the bijection between the nodes is also easy to infer.



(a)

Motif signature:

$\{(\{\},\{0\}),(\{0,1\},\{\}),(\{\},\{1\})\}$

Orbit signature:

$(\{0,1\},\{\})$

(b)

Motif signature:

$\{(\{\},\{0\}),(\{0,1\},\{\}),(\{\},\{1\})\}$

Orbit signature:

$(\{0,1\},\{\})$

Figure 2: A and D occupy the same automorphism orbit in these motifs because the motif signatures and the orbit signatures both match.

For any given temporal graph, we can compute a signature of the graph which is guaranteed to be equal to the signature of another graph if and only if they are isomorphic.

1. We first sort the edges by time and label them with their index in the sorted order.
2. We compute the signature of a node as a 2-tuple; the set of indices of all outgoing edges and all incoming edges.
3. The set of signatures for each node in the graph provides a signature for the motif.

The signature of the node under study identifies the automorphism orbit within the motif.

4.3 Motif Sampling

The number of motifs in a graph can grow rapidly with the average degree of the nodes in the graph. This makes it difficult to compute motif degree vectors for dense graphs or graphs with a power law degree distribution. However, it is possible to sample from the set of all motifs to approximate the true distribution. To do so, modify the earlier subgraph enumeration algorithm, but instead of extending the motif with each edge in the extension set, pick one uniformly at random (discarding all earlier edges). This will sample up to one motif non-uniformly at random. The probability of a motif i being sampled is the product of the sizes of the extension sets from which the edges in the motif are sampled. Let $E_{ext}^{(j)}$ be the extension set from which the edge j of motif i is sampled from.

$$p_i = \prod_{j=1}^{k} \frac{1}{|E_{ext}^{(j)}|}$$

To ensure that for every motif adjacent to the node of study, the expected value of the update is 1, we increment the motif's score by $\frac{1}{p_i}$. Continue sampling motif instances in this fashion until the variance is as small as desired.

5 Results

5.1 Guaranteed Convergence for Motif Sampling

Let X be a random variable corresponding to the update to a particular motif, where the local graph contains n instances of the motif with sampling probabilities $p_1, p_2, ...p_n$.

Theorem 1. X is an unbiased estimator

$$E[X] = n$$

Proof. The estimator will be incremented by $\frac{1}{p_i}$ with probability p_i if motif i is counted.

$$X = \sum_{i=1}^{n} \frac{1}{p_i} \text{Bernoulli}(p_i)$$

$$E[X] = \sum_{i=1}^{n} \frac{p_i}{p_i} = n$$

□

We can also show that the estimator converges favorably.

Theorem 2. Let X_s be the mean of s independent samples from X. If we select at least $\frac{k!D^k}{\alpha^2 n}$ samples, it is guaranteed that X_s will have standard deviation less than αn. D is the maximum degree of the graph, and k is the number of edges in the motifs.

Proof.

$$Var(X) = E[X^2] - E[X]^2$$

Any two distinct motifs, $i \neq j$ are sample independently, so the probability of both being sampled is Bernoulli$(p_i p_j)$. Since the probability of sampling motif i is entirely correlated with itself, that term of the sum has probability Bernoulli(p_i).

$$Var(X) = E[(\sum_{i \neq j \leq n} \frac{1}{p_i p_j} \text{Bernoulli}(p_i p_j))^2 + \sum_{i=1}^{n} \frac{1}{p_i^2} \text{Bernoulli}(p_i)] - n^2$$

$$Var(X) = \sum_{i \neq j \leq n} 1 + \sum_{i=1}^{n} \frac{1}{p_i} - n^2$$

$$Var(X) = (n^2 - n) + \sum_{i=1}^{n} \frac{1}{p_i} - n^2$$

$$Var(X) = \sum_{i=1}^{n} \frac{1}{p_i} - n$$

$$\frac{1}{p_i} = \prod_{j=1}^{k} |E_{ext}^{(j)}|$$

$E_{subgraph}^{(j)}$ contains $j-1$ edges, and therefore it contains at most j nodes. All edges in $E_{ext}^{(j)}$ must be adjacent to a node in $E_{subgraph}^{(j)}$, so $|E_{ext}^{(j)}| \leq jD$ where D is the maximum degree of the graph.

$$\frac{1}{p_i} \leq \prod_{j=1}^{k} jD = k!D^k$$

$$Var(X) \leq nk!D^k$$

$$\sqrt{Var(X_s)} = \sqrt{\frac{Var(X)}{s}} \leq \sqrt{\frac{nk!D^k}{s}}$$

$$\sqrt{\frac{nk!D^k}{s}} \leq \alpha n$$

$$s \geq \frac{k!D^k}{\alpha^2 n}$$

□

If you uniformly increase the density of a temporal graph, the number of samples required to achieve a fixed motif estimate error α is fixed.

Theorem 3. Let $G = (V, E)$ be a randomly generated temporal graph for which for every pair of nodes $u, v \in V$ the probability density of the edge (u, v, t) existing for all $t \in [0, T]$ is equal to βA_{uv} and these probabilities are independent. Then then number of samples s required to estimate the motif vector is independent of β.

Proof. The expected maximum degree D of G scales proportionally to

$$T\beta \max_{u \in V}(\sum_{v \in V} A_{uv} + A_{vu}) = O(\beta)$$

There are some additional constant factors, but they tend to 1 as $\beta \to \infty$, and they are generally not relevant for the proof.

The expected number of instances is equal to the integral over all the potential ways that motif could appear, weighted by the probability of each of these instances actually appearing. For instance, for the reciprocated edge motif, you need two edges $(u, v, t_1), (v, u, t_2)$. The set of potential instances of this motif over a set of edges V and an interval of time $[0, T]$ is independent of the actual structure of the graph. The probability of each of these motifs (because we assume edge probabilities are independent) is the product of the probability density of the existence of any given edge, proportional to $A_{uv}A_{vu}$. For any motif, the expected number of instances scales as $O(\beta^k)$ where k is the number of edges in the motif.

The required number of samples proven in theorem 2, $\frac{k!D^k}{\alpha^2 n}$, scales as $\frac{O(\beta)^k}{O(\beta^k)} = O(1)$. Therefore the number of samples required is independent of the density of the graph, assuming the probability of edges appearing in the graph are independent. □

5.2 Financial Transaction Data Set

The network we are studying consists of financial transactions over 50,000 euros in a small European country between 2008 and 2015. The graph includes 118,739 companies, including banks, government agencies, and individuals which constitute the nodes of the graph. The temporal edges are 2,982,049 transactions over these 8 years. The companies are also labeled with additional information, such as what industry they are part of and yearly balance sheet information.

5.3 Motif Analysis

To test for differing network structure between different subpopulations within this transaction graph, we first sample 1000 nodes uniformly at random, then sample 100 motifs for each node using the above sampled motif count estimator. For the purposes of this analysis, we set $\delta = 30$ days and look for motifs with 2 edges. One could count larger motifs, but the number of distinct motifs grows exponentially with the number of edges. There are few enough 2 edge motifs to visualize them and comment on them manually. Larger motifs would be useful for generating more expressive feature vectors for nodes in a graph.

Nodes in the graph have drastically different degrees, and we would like to make sure any differences we see are not merely consequences of degree. Therefore, we divide the motif count of any given motif by the degree of the studied node to the power of the number of edges in the motif which are adjacent to the studied node. Assuming edges are uniformly distributed through time and direction, this eliminates the effect of the degree of the ego node from the resulting motif vector. We then normalize each vector so all the elements for a given node sum to 1.

To test if a subpopulation has a different structure than typical nodes in the graph, we use the Anderson-Darling statistical test to compare the distribution of each element in the normalized motif vector to the distribution in the population as a whole. In some cases, there is a significant difference in the distribution, but only a small difference in the means of the distributions. To see a comparison of the motif vectors of various subpopulations, see figure 1 in the Appendix.

Motifs in which the studied node is at the edge, rather than the center, are far more prominent. This is likely a consequence of the dissortativity of this graph (R=-0.258). The combination of a power law degree distribution and dissortative connectivity implies that most nodes are low degree, but often have high degree neighbors. Therefore there are few motifs where both edges are adjacent to ego, but many which go two hops away.

6 Conclusion

Here we have presented a novel method for measuring micro-scale structure of temporal graphs. This is a generalization of earlier work by Paranjape et. al, Wernicke et. al, etc. We also introduce a technique for sampling temporal motifs to estimate their densities far more efficiently than full enumeration. We apply this technique to a real world graph, and observe that the temporal structure of different nodes vary based on other out-of-graph attributes of those nodes. Further work can derive more value from these temporal graph motif vectors by potentially using that for predictive, rather than exploratory, tasks.

7 Team Member Contributions

I am a one person team, and therefore did everything.

References

[1] Ashwin Paranjape and Austin R. Benson and Jure Leskovec. *Motifs in Temporal Networks.* International Conference on Web Search and Data Mining, 2017.

[2] Baharan Mirzasoleiman. *Analytical Null Models for Temporal Motifs.* [*Discovering Trends and Anomalies in Dynamic Networks*]. In Proceedings of ACM WSDM conference (WSDM2019). ACM, New York, NY, USA, 9 pages.

[3] Lauri Kovanen, Marton Karsai, Kimmo Kaski1, Janos, Kertesz, and Jari Saramaki, *Temporal motifs in time-dependent networks.* J. Stat. Mech. 2011.

[4] Sebastian Wernicke, Florian Rasche. *FANMOD: a tool for fast network motif detection*, Bioinformatics, Volume 22, Issue 9, 1 May 2006, Pages 1152–1153, https://doi.org/10.1093/bioinformatics/btl038

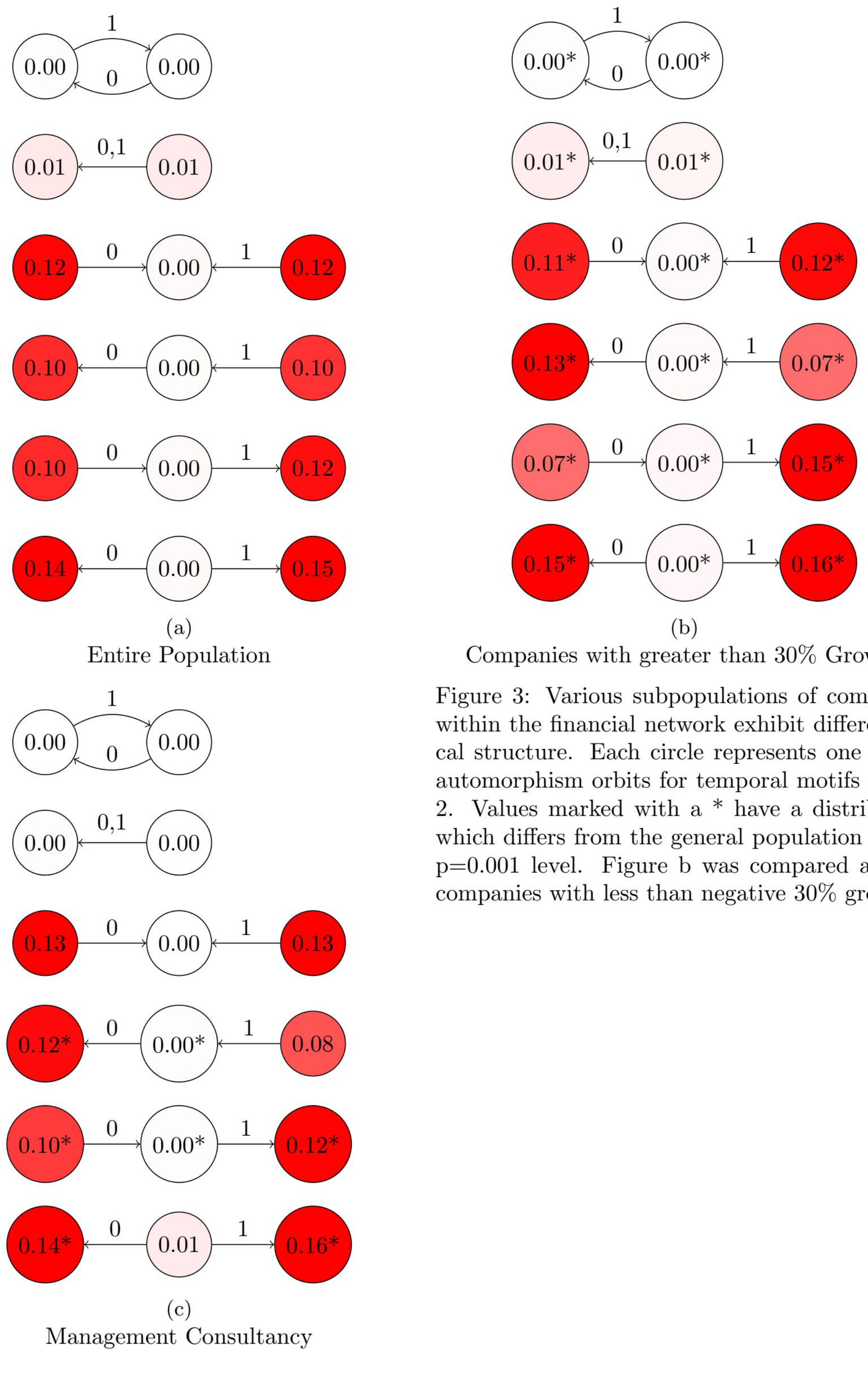
[5] Tijana Milenković and Nataša Przulj, *Uncovering Biological Network Function via Graphlet Degree Signatures.* Schedule of the RECOMB Satellite Conference on Systems Biology. 2007.

[6] Yael Artzy-Randrup, Sarel Fleishman, Nir Ben-Tal, and Lewi Stone, Comment on ”Network Motifs: Simple Building Blocks of Complex Networks” and ”Superfamilies of Evolved and Designed Networks”, Science, 2004.

8 Appendix

Source code available at

```
https://github.com/benhannel/cs224w_temporal_motifs
```



(a)
Entire Population

(b)
Companies with greater than 30% Growth

(c)
Management Consultancy

Figure 3: Various subpopulations of companies within the financial network exhibit different local structure. Each circle represents one of the automorphism orbits for temporal motifs of size 2. Values marked with a * have a distribution which differs from the general population at the p=0.001 level. Figure b was compared against companies with less than negative 30% growth.