Network Neighborhood Analysis

Michael D. Porter
SPADAC
McLean, VA 22102
Email: mike.porter@spadac.com

Ryan Smith
SPADAC
McLean, VA 22102
Email: ryan.smith@spadac.com

***Abstract*—We present a technique to represent the structure of large social networks through ego-centered network neighborhoods. This provides a local view of the network, focusing on the vertices and their k^{th} order neighborhoods allowing discovery of interesting patterns and features of the network that would be hidden in a global network analysis. We present several examples from a corporate phone call network revealing the ability of our methods to discover interesting network behavior that is only available at the local level. In addition, we present an approach to use these concepts to identify abrupt or subtle anomalies in dynamic networks.**

I. Introduction

There are numerous ways to represent the structure of a social network. However, with the size and complexity of modern dynamic networks, finding the appropriate representation to convey the desired network information is a challenging task. Besides the computational difficulties, large networks also complicate analysis as visualization and inference becomes increasingly impractical for the full network.

However, like any large data set, summary statistics (e.g. graph invariants) are one way to help succinctly describe certain aspects of the networks. Another approach is to break up the network into smaller, easier to manage components and study the properties of the sub-networks. The approach taken in this paper is to combine these two concepts for the purpose of analyzing large temporally varying social networks.

This is accomplished by considering certain vertex or subgraph level locality statistics specified on local regions of a network. These local regions are defined as the neighborhoods around the vertices (i.e. ego networks). Neighborhood analysis can reveal certain aspects of the network that are concealed when only aggregate global network measures are considered. This allows the small patterns, anomalies, and features (as might be relevant to crime and terrorism networks) to be discovered that would be missed in a more global analysis. For example, identifying all the local leadership changes or increased activity regions can help identify terrorist cells [1].

We consider network neighborhood analysis for two scenarios: identifying interesting features of a network snapshot (at a specific point in time) and detecting anomalies in a dynamic network over time. The choice of summary statistic can provide different information about the network and detection of certain types of changes. This is a very general and flexible set-up that can be scaled to consider single vertex metrics to full graph measures. This also enables a focus to be given to smaller regions of the network where visualization and inference tools can be used more effectively and efficiently.

II. Formulation

We consider time indexed undirected binary graphs $G_t = (V_t, E_t)$, where V_t is the set of N_t vertices and E_t are the M_t edges at time t. Such graphs can be fully represented by an adjacency matrix $A_t = \{A_t(i,j)\}_{i,j \in V_t}$ where $A_t(i,j) = 1$ if $(i,j) \in E_t$ (i.e. an edge connects vertices i and j) and 0 if there is no edge between them.

A. Neighborhood Definition

The goal is to break up a large complex network into smaller local regions that can better reveal interesting features and structure. The approach taken here is to use the family of vertex anchored neighborhoods (often called ego networks [2]) as our unit of analysis. This defines a class of local sub-networks that can help identify vertices, or communities of vertices, that share interesting features.

Formally, let $d_t(u,v) \in \{0,1,\ldots\}$ be the shortest path distance (geodesic) between the vertices u and v at time t. A vertex's k^{th} - order neighborhood is defined as the set of vertices within a distance of k. Formally, $N_t[v,k] = \{u \in V_t : d(u,v) \leq k\}$ is the k^{th} order closed neighborhood (set of vertices) around, and including, vertex v and time t.

The neighborhood subgraph $g(N_t[v,k])$ is the subgraph induced by a neighborhood. It consists of the vertices: $N_t[v,k]$ and edges: $\{e_t(i,j) \in E_t : i,j \in N_t[v,k]\}$. Let $n_t(v,k) = |N_t[v,k]|$ be the number of vertices and $m_t(v,k) = |\{e_t(i,j) \in E_t : i,j \in N_t[v,k]\}|$ be the number of edges in the neighborhood. This neighborhood definition creates a family of subgraphs $\mathcal{G}_t = \{g(N_t[v,k]) : v \in V_t, k = 0,1,\ldots\}$ for each time period. Figure 1 gives an example of a set of vertex anchored neighborhoods.

For each subgraph, we calculate a neighborhood statistic $S_t(v,k) \equiv S(g(N_t[v,k]))$, describing the k^{th} order neighborhood around vertex v. The choice of statistic will provide certain details of the vertex and it's neighborhood.

It is important to recognize that this set of subgraphs is neither exhaustive nor disjoint. These neighborhoods, while large, are not the only way to define local regions of the network. However, they are computationally convenient and in the lack of other information, a natural definition of vertex locality.

978-1-4244-6446-3/10/$26.00 © 2010 IEEE

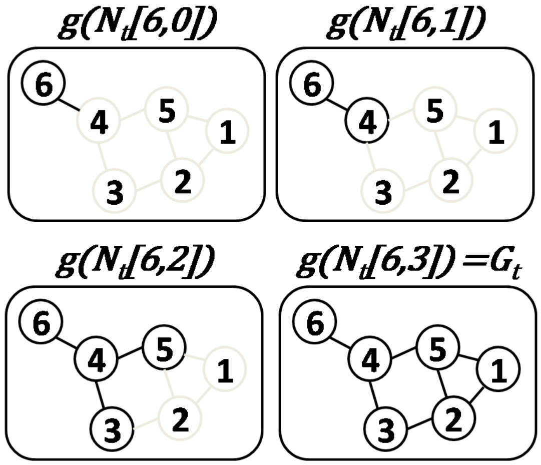

Fig. 1. The darkened edges and vertices correspond to the elements in the network neighborhood anchored at vertex 6. The 3^{rd} order neighborhood (i.e. $k = 3$) corresponds to the full network.

The decomposition of a network into neighborhoods has several advantages. First, the smaller order neighborhood can reveal important information about the individual vertices' "position" in the network [3], [4] allowing an analyst to quickly identify the vertices that satisfy a certain position (e.g. broker, leader, bridge). In addition, the neighborhoods also specify a community from which subgraph level statistics can be ascertained. Community characteristics, like density, can be used to identify tightly coupled regions of the network.

B. *Neighborhood Metrics*

There are numerous metrics that can be calculated for networks. Some are specific to vertices (e.g. degree) and others describe an aspect of the entire network (e.g. density). However, all metrics are dependent on the specification of the network and their values can change (sometimes drastically) if the network composition changes. This makes the specification of the network a very important task, and one that is often crucial to the success of network analysis [5]. To minimize the effects of network selection, the neighborhood representation will allow evaluation of smaller parts of the network, allowing smaller scale effects to be captured, in addition to being sufficiently large to also capture the large scale effects.

For a given network graph G_t we define a set of metrics $S_t(v, k)$ measuring some aspect of the subgraph $N_t[v, k]$ for every vertex v over a set of sizes ($k = 0, 1, 2, \ldots$). When k equals the diameter of G_t, then all neighborhood metrics converge to the global network value. Thus, our neighborhood definition facilitates a multi-scale analysis of the network. When vertex level inference is desired, the neighborhood size is usually limited to $k = 3$, however for community level inference, the network size can be increased to the full global network.

Our premise is that interesting features of a network can appear locally but be hidden with metrics based on the entire network. For example, betweenness centrality of a vertex is a measure of it's influence in a network [6]. It is calculated from the number of shortest paths (geodesics) between all vertices in a network that traverse vertex v. As a global measure, this will be effected by all the other vertices in a network, even those that are distant. However, as a local view of influence, betweenness centrality calculated in the neighborhoods can provide a different interpretation. For the neighborhood measure, the centrality of vertex v is only calculated from the other vertices in their neighborhood. This estimates the local effect a vertex has and can be a valuable way to identify locally important vertices.

In addition to focusing on the vertices, we can use neighborhood analysis to identify subgraphs that have interesting community characteristics. For example, the density of a network is the ratio of the number of edges to the number of possible edges. This gives a measure of how cohesive or close a network is. A global measure will average out the local effects and give no insight into the local network dynamics. Alternatively, by defining the density at the neighborhood level $dens(N_t[V, k]) = 2 \cdot m_t(v, k)/\left(n_t(v, k)(n_t(v, k) - 1)\right)$, a more detailed picture of the network can now be discovered. Neighborhoods with $dens(N_t[V, k]) = 1$ are cliques.

III. Exploring Networks over Vertices and Size

Using internal company phone records, we can demonstrate the usefulness of our methodology. The data consists of the times of calls between SPADAC company employees between May 2008 - May 2009. We only consider weekly undirected graphs in this analysis creating edges between two individuals if they had a conversation (of any duration) during the week. Figure 2 shows some basic metrics for the entire network.

Consider the network snapshot (i.e. time is fixed at t), G_t. We can view the statistics of the neighborhood by vertex and size. For example, Figure 3 shows a view of (standardized) betweenness centrality for all vertices across neighborhood size. Notice that not all the vertices with a large (small) global betweenness centrality also have large (small) local betweenness. This is contradictory to the analysis in [7] who evaluated random graphs only. For the same time period, Figure 4 shows the density statistic and Figure 5 gives the subgraphs for the two noted vertices. In general, we will calculate a set of statistics for each neighborhood in this manner to determine the local structure of the network.

Vertex 169 is in the position of a bridge or broker connecting two larger groups. Alternatively, vertex 157 may be in a position of a local leader, connecting multiple smaller groups. While the vertices both have high betweenness at the local level, they are not seen as important at the global level.

IV. Exploring dynamic networks

In addition to using the neighborhood formulation to examine interesting regions of a network at a snapshot in time, it can also be useful for exploring the network dynamics as time evolves. In particular, we examine these local neighborhood regions over time to detect anomalies. These can correspond to a vertex that undergoes a change in status (e.g. takes leadership

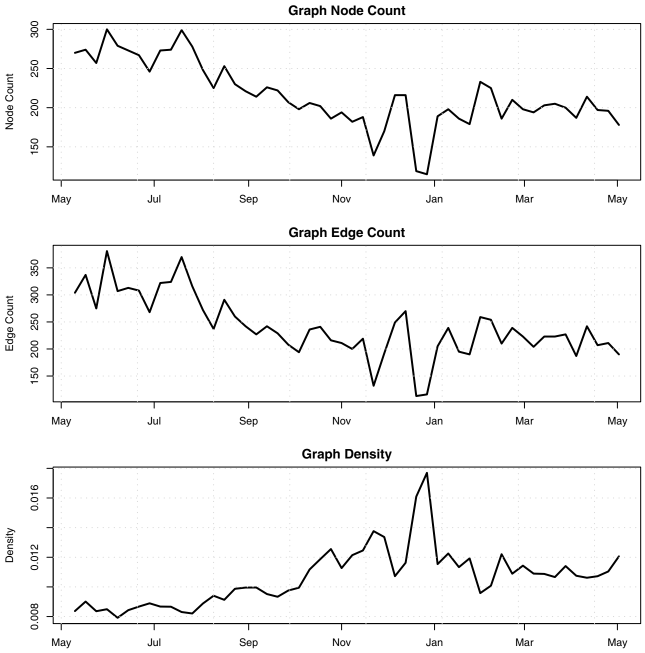

Fig. 2. Summary statistics for the weekly networks from company phone records for 2008-2009. The large jumps correspond to the Thanksgiving and Christmas holidays. The density appears to increase around August corresponding to a company restructuring.

role, gets demoted) or a community that changes function (e.g. terrorist cell moves from sleeper to planning); changes that could be missed if a full network was examined.

A. Anomaly Detection

For detecting anomalies in networks, we adopt an approach similar to the network scan statistics adopted by [8], [9]. That is, we calculate a discrepancy measure for each neighborhood (over all sizes). The discrepancy is a measure of the unusualness of an observation.

Based on the neighborhood statistic employed, we can define a discrepancy measure $D_t(B)$ describing how unusual the subgraph, given by the vertex and neighborhood size, $B = (v, k)$ appears at time t. These measures should be suitably standardized to allow direct comparison between all neighborhood sizes and times. The scan statistic of the graph at time t is the maximum discrepancy over all neighborhoods and sizes

$$M_t = \max_{B \in \mathcal{B}} D_t(B)$$

where $\mathcal{B} = \{(v, k) : v \in V, k = 0, 1, 2, \ldots\}$ and V is the set of all possible vertices.

However, unless there are very large abrupt changes, this may not detect the change with sufficient power. Therefore, we extend the discrepancy measure to accumulate the evidence of change from time s to t with a time indexed scan statistic

$$M_{t,s} = \max_{B \in \mathcal{B}} D_t(B, s)$$

where $D_t(B, s)$ is the discrepancy of the observations in B in the time period $[s, t]$. The B that obtains the maximum value

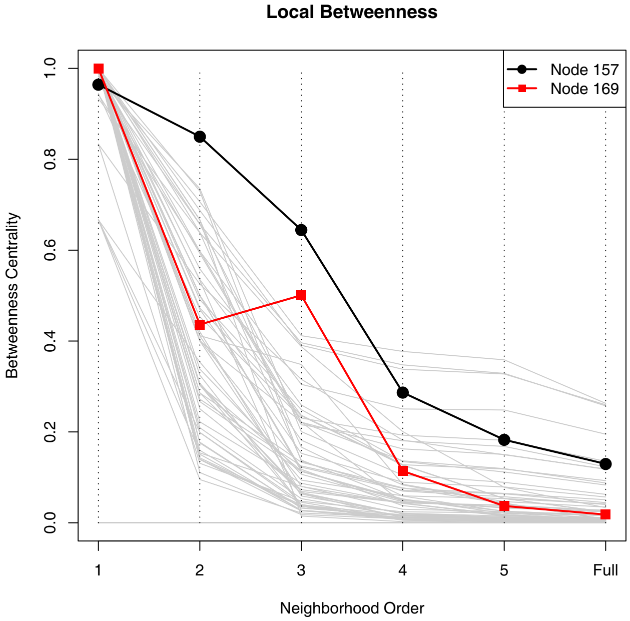

Fig. 3. Betweenness centrality for $t = 40$. The gray lines show the centrality for all vertices across the neighborhood size. Two particular vertices are highlighted, each corresponding to an unusual observation.

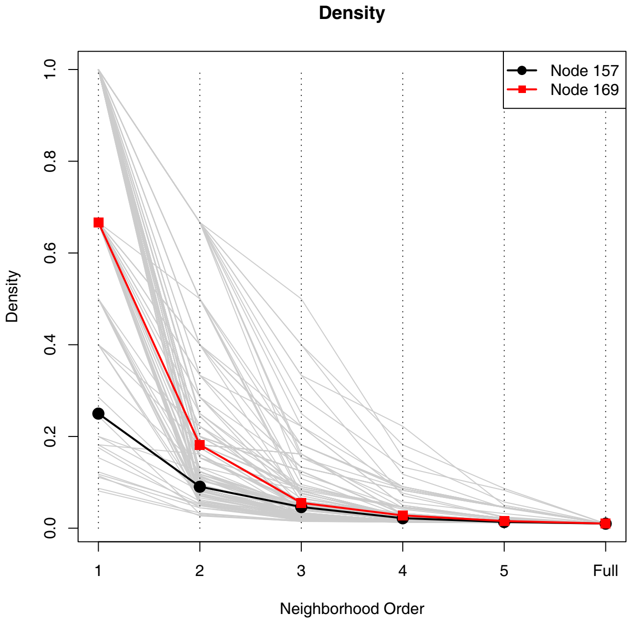

Fig. 4. Graph density for $t = 40$. The colored paths correspond to the same vertices in Figure 3, showing how the two different metrics provide different views of the network structure.

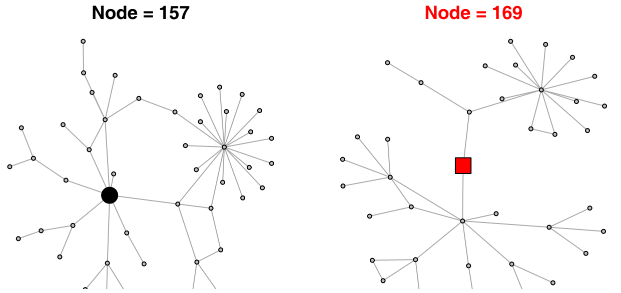

Fig. 5. The subgraph for vertices $v = 157, 169$ for neighborhood size $k = 3$ at time $t = 40$. These two vertices had the largest betweenness scores of any 3^{rd} order neighborhood.

provides the vertex and neighborhood size that produce the most anomalous results given that a change started at time s. For signaling an alarm, we use Lorden's GLR statistic [10] as the overall change measure

$$R_t = \max_s M_{t,s}$$

The s that provides the maximum is an estimate of the change point.

B. Discrepancy Measure

A common discrepancy measure, and one we advocate here, is the likelihood ratio [11]. This requires a distributional assumption on the observed values of the statistics. While social networks do not usually possess common distributions, by defining discrepancy locally and keeping the neighborhood fixed through time such assumptions become more plausible.

For example, the density statistic is the ratio of edges in a neighborhood to the total possible number of edges. It may be plausible to assume this follows a binomial distribution with parameter p at any point in time providing the likelihood ratio discrepancy

$$D_t(B,s) = \prod_{i=s}^{t} \left(\frac{\hat{p}_1}{\hat{p}_0}\right)^{x_i} \left(\frac{1-\hat{p}_1}{1-\hat{p}_0}\right)^{X_i - x_i}$$

where $x_i = m_i(B)$ is the number of edges and $X_i = n_i(B)[n_i(B)-1]/2$ the number of possible edges in neighborhood B at time i. Defining $p_{u:v} = \sum_{i=u}^{v} x_i / \sum_{i=u}^{v} X_i$ the maximum likelihood estimates for p are $\hat{p}_1 = p_{s:t}$ (assuming change started at time s) and $\hat{p}_0 = p_{1:t}$ (assuming no change).

Several other statistics, when standardized, have the same form. For example, betweenness centrality can be standardized by defining it as the number of geodesics that traverse a vertex divided by the total number of geodesics in the network. In this case, we would take $x_i = \delta_i(v,B)$, the number of shortest paths in the neighborhood B that traverse the anchor vertex v and $X_i = [n_i(B)-1][n_i(B)-2]$, the number of shortest paths in the neighborhood (excluding those originating or terminating at v).

Figure 6 shows the time indexed scan statistics for density by neighborhood size $M_{t,s}(k) = \max_{v \in V} D_t((v,k),s)$.

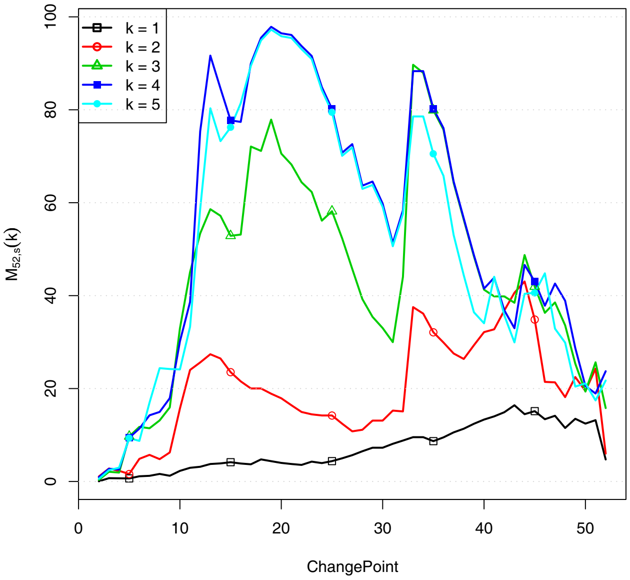

Fig. 6. Time indexed scan statistics for density by neighborhood size.

The greatest discrepancy occurs in the largest neighborhood sizes ($k = 3, 4, 5$) around $t = 20$ (late-August), the time corresponding to the company restructuring. While it may be interesting to investigate what vertices and communities lead to this large change, we instead focus on other changes that may have occurred in the network.

By taking a more local view ($k = 1, 2$) the most likely change point appears to be around time $t = 44$ (early-March). This is something not revealed in the global statistics from Figure 2. The density statistic and discrepancy over time for the neighborhood given by $B = (8, 2)$, the vertex providing the largest discrepancy for a neighborhood size of $k = 2$ is shown in Figure 7. The discrepancy measures shows the most likely change point occurring at $s = 44$ (given we have received observations up to $t = 52$). A graph view of the second order neighborhoods for vertex $v = 8$ is plotted for all time periods in Figure 8. It appears from this plot that the density dropped as the vertex became better connected.

V. Conclusion

We described a methodology for network analysis based on the concept of local network neighborhoods. This allows analysis on large social networks and can reveal aspects of dynamic network structure that cannot be ascertained in a global network analysis. We provided some examples from a company phone network displaying the flexibility of our method to visualize and discover interesting patterns by vertex, neighborhood size, and time. In addition, we described an anomaly detection methodology with the ability to discover anomalous vertices and communities that are hidden to standard analyses.

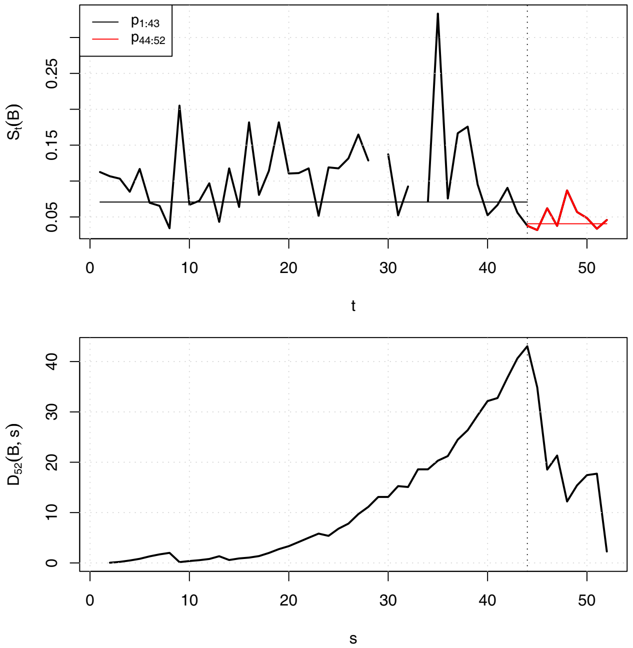

Fig. 7. The top shows the density statistic for $B = (8, 2)$ (i.e. vertex $v = 8$ and size $k = 2$) along with the estimates of the density in the two time periods. The bottom plots the discrepancy. The most likely change point is $t = 44$, corresponding to the time when the density dropped the most significantly.

References

[1] V. E. Krebs, “Mapping networks of terrorist cells,” *Connections*, vol. 24, no. 3, pp. 43 – 52, 2002.

[2] R. A. Hanneman and M. Riddle, *Introduction to social network methods*. Riverside, CA: University of California, Riverside, 2005. [Online]. Available: http://faculty.ucr.edu/h̃anneman/

[3] S. P. Borgatti and M. Everett, “Notions of position in social network analysis,” *Sociological Methodology*, vol. 22, no. 1, pp. 1 – 35, 1992.

[4] R. S. Burt, “Structural holes and good ideas,” *American Journal of Sociology*, vol. 110, no. 2, pp. 349 – 399, 2004.

[5] S. Wasserman and K. Faust, *Social network analysis: methods and applications*. Cambridge: Cambridge University Press, 1994.

[6] U. Brandes, “A faster algorithm for betweenness centrality,” *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163 – 177, 2001.

[7] M. Everett and S. P. Borgatti, “Ego network betweenness,” *Social Networks*, vol. 27, no. 1, pp. 31 – 38, 2005.

[8] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, “Scan statistics on enron graphs,” *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 229 – 247, 2005.

[9] C. Priebe and W. Wallis, “On the anomalous behaviour of a class of locality statistics,” *Discrete Mathematics*, vol. 308, no. 10, pp. 2034 – 2037, 2008.

[10] G. Lorden, “Procedures for reacting to a change in distribution,” *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[11] M. Frisén, “Statistical surveillance. optimality and methods,” *International Statistical Review*, vol. 71, no. 2, pp. 403–434, 2003.

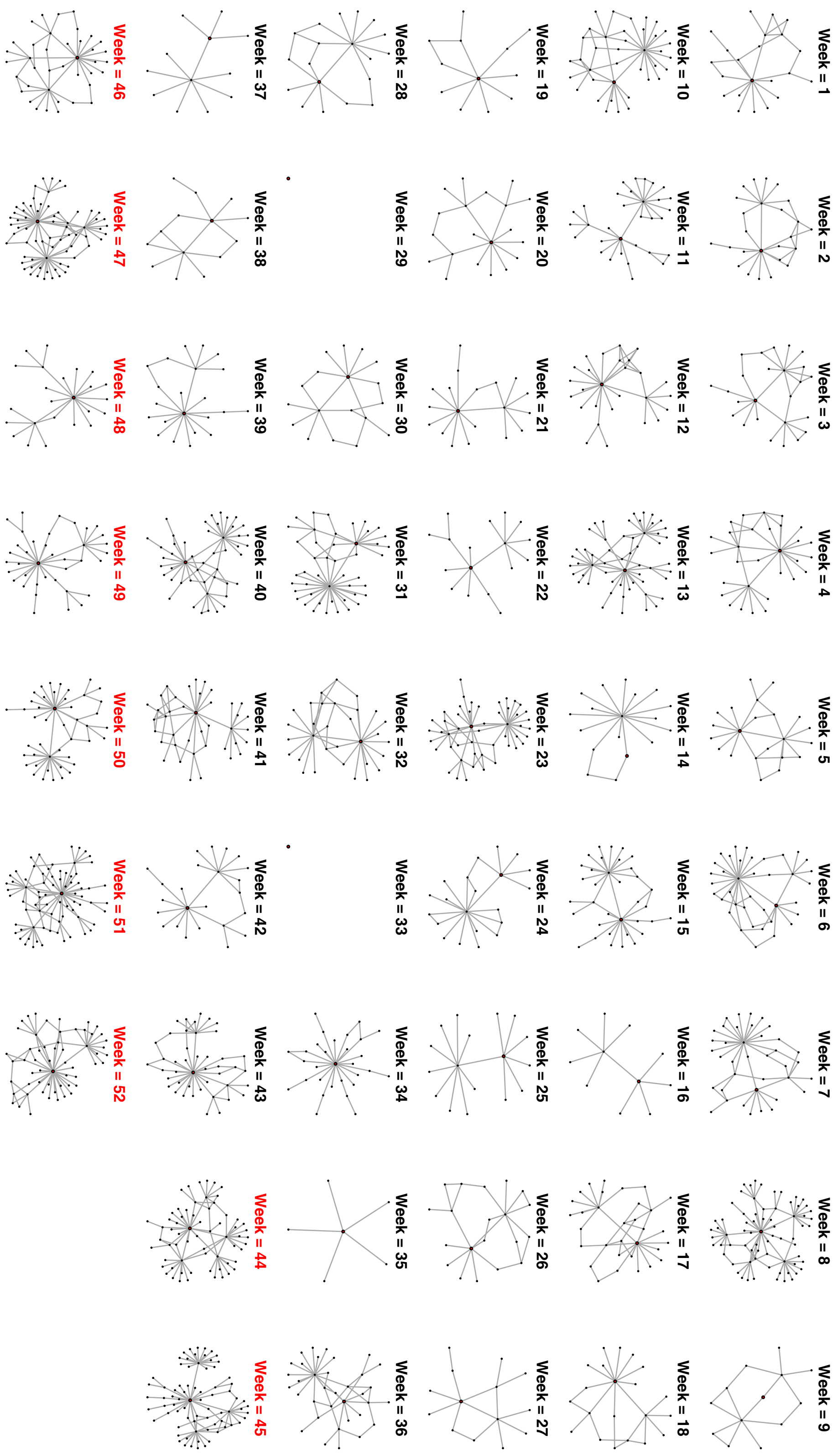

Fig. 8. Graph view of local neighborhood over time for the vertex with highest discrepancy for size $k = 2$.