

An Ensemble Framework for Detecting Community Changes in Dynamic Networks

Timothy La Fond, Geoffrey Sanders, Christine Klymko, and Van Emden Henson

Lawrence Livermore National Laboratory

Livermore, California 94550

Abstract—Dynamic networks, especially those representing social networks, undergo constant evolution of their community structure over time. Nodes can migrate between different communities, communities can split into multiple new communities, communities can merge together, etc. In order to represent dynamic networks with evolving communities it is essential to use a dynamic model rather than a static one. Here we use a dynamic stochastic block model where the underlying block model is different at different times. In order to represent the structural changes expressed by this dynamic model the network will be split into discrete time segments and a clustering algorithm will assign block memberships for each segment. In this paper we show that using an ensemble of clustering assignments accommodates for the variance in scalable clustering algorithms and produces superior results in terms of pairwise-precision and pairwise-recall. We also demonstrate that the dynamic clustering produced by the ensemble can be visualized as a flowchart which encapsulates the community evolution succinctly.

1. Introduction

Over the past decade or so, network analysis has emerged as an important area of mathematics and computer science, as well as in many other disciplines (see [1], [2], [3] among many others). One of the significant questions in network analysis is how to cluster the nodes in the network to identify underlying communities [4], [5]. Clustering static graphs reveals important community structure that can be leveraged for many important analysis tasks involving relational datasets. Several fast clustering techniques have been defined for detecting communities that have high internal connectivity relative to their external connectivity [6], [7], [8], [9]. Analysis of community structure in *dynamic graphs*, where each edge is associated with a timestamp, is vastly more complex. Many real-world networks experience concept drift as the networks evolve; such evolutions cannot be accurately represented with a static model. Ignoring the temporal information and applying static clustering techniques entire dataset risks missing important structural changes; important but transient community development could remain undiscovered.

Several types of changes in community structure are of high analytic interest, including (i) the birth/death of a community [10], (ii) the growth/decay of a community [10], (iii) the changes of internal connectivity of a community

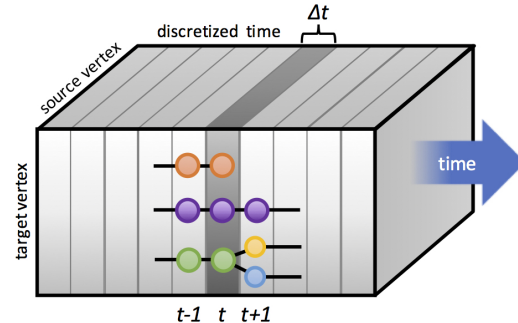


Figure 1: Schematic of the dynamic community detection framework employed. For each timestep, t , a representative partitioning is determined. Representative partitionings from adjacent timesteps ($t - 1$) and $t + 1$ are compared for tracking communities and their dynamic structural changes such as community death (dark orange) or community splitting (green into blue and yellow).

(such as a dense community becoming increasingly bipartite [11]), (iv) groups of communities splitting/merging [10], [12], and (v) member migration between groups of two or more communities (as well as the identification of nodes that do not migrate [13]). Clearly, there is a need for both the ability to perform *dynamic community detection* to find the community behavior at any given time and to *represent community evolution* to display the kinds of temporal effects that occur in the network. In this paper, we design a framework for taking a static community detection algorithm and employing it to find the evolving community structure of dynamic networks. Our framework also provides an interpretable visualization of the major changes to the community structure.

Some dynamic community detection algorithms, for example, GraphScope, have memory and will continue to use the discovered data model from prior time steps as long as that model passes some performance metrics [11]. This memory effect can sometimes be detrimental if the algorithm fails to update the model during a change in the underlying structure. Other algorithms have no memory at all and merely connect the results of independent community detection passes run at different times in the network [10]. We will describe how to apply our framework in both a memory and memoryless fashion, where the version with memory attempts to smooth the temporal noise of the data at the cost of sometimes lagging behind immediate changes

to the underlying model.

Often edge timestamps are *essentially continuous*, where the resolution of timestamps is on much smaller intervals than the span of time required to resolve communities, both statically and during their evolution. For example, we might have timestamps that have resolution in the seconds while a few hours of connectivity data are necessary to resolve most community structure present in the dataset. A large class of dynamic graph analysis techniques discretize the time dimension into a number of uniform non-overlapping *time segments*.

The selection of an appropriate timestep width which provides sufficient resolution to detect temporal events remains an open problem, with time segments typically being defined to start and end at a natural frequency determined by the data source, e.g. daily or weekly. We will also restrict ourselves to community dynamics (i) and (iv) (birth, death, splitting, and merging).

In this paper we:

- Define a dynamic stochastic block model capable of producing data with community evolution events.
- Introduce a *Dynamic Clustering Ensemble Framework* which, given a baseline static clustering algorithm, finds the dynamic community behavior of a network and reports any evolution in the community structure.
- Demonstrate that the ensemble framework has superior pairwise-precision and pairwise-recall performance in recovering block assignments than the baseline algorithm. We apply this ensemble to synthetic data, drawn from both static and dynamic models, as well as to semi-synthetic data.
- Create a visualization of the community evolution in a real world dataset and illustrate the types of community dynamics found (Figure 7).

2. Approach

2.1. Data Model

To aid in the development of temporal algorithms, generative dynamic graph models are often used. These models support: (a) building fully synthetic examples with various community dynamics, allowing algorithm validation on examples with strong ground-truth; (b) injecting synthetic community dynamics into real world dynamic graphs, allowing algorithm validation in the face of realistic noise; and (c) gauging the statistical significance of various community events that have been detected. Degree-corrected stochastic block models whose model parameters are functions of continuous time are able to exhibit the kinds of phenomenon described in (i)-(v) above, and we will be using a version for our synthetic and semi-synthetic data analysis.

In order to create the type of community dynamics we expect to see in real networks we introduce a dynamic stochastic block model as a generative graph model where

every edge A_{ijt} is sampled according to a Poisson distribution with a time-dependent rate λ_{ij} :

$$A_{ijt} \sim \text{Poisson}(\lambda_{ij}(t)) \quad \lambda_{ij}(t) = \theta_i \theta_j \Sigma_{b_i, b_j}(t)$$

where θ_i controls node i 's expected degree, b_i is i 's block assignment, and $\Sigma_{b_i, b_j}(t)$ is the interaction strength between blocks b_i and b_j at time t . As the block model is a function of time, by gradually changing the edge probability between collections of nodes over time we can introduce smooth splits and merges into the community structure. For example, if the value of $\Sigma_{b_i, b_j}(t)$ grows linearly with increasing t we effectively cause b_i and b_j to merge into a single block as their interconnectivity grows.

2.2. Ensemble Framework

Given a dynamic graph as defined above and an *a priori* division of the graph stream into discrete time segments, the algorithm presented clusters the graph at every segment, using as a baseline the Markov chain Monte Carlo (MCMC) algorithm described in the works of Peixoto [7], [8], [9]. This algorithm uses an ergodic Markov chain to randomly modify the block membership of each node and probabilistically accepts or rejects each modification. Although the algorithm will converge given a sufficiently long mixing time, the required mixing times are often prohibitively long and, in practice, approximations to the MCMC process are used. These approximations, combined with the non-determinism of the MCMC process, lead to variability in the produced clustering given different runs of the MCMC algorithm on the same data.

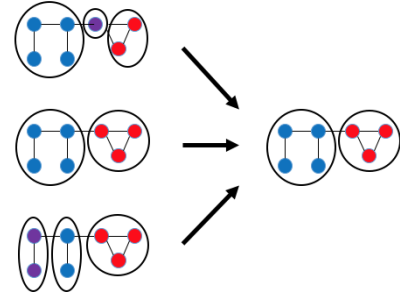


Figure 2: The clusterings determined by various runs of the MCMC clustering algorithm are used to form a representative clustering of the nodes.

We address this variability using an ensemble approach which performs multiple runs of the MCMC clustering algorithm at each time segment. This accommodates for variance in the clustering algorithm by applying the algorithm multiple times and obtaining a cloud of possible partitions. These clouds are then resolved to produce a best overall cluster assignment for the time segment, referred to as the *representative clustering*, which corresponds to the most typical partitions seen in the cloud. A small example of this can be seen in Figure 2.

Let $H(X, Y)$ be any similarity function across two sets of nodes. Given K runs of the MCMC clustering algorithm at time t , leading to K different partitions of the form $\{B_1^{k,t}, B_2^{k,t}, \dots, B_{c_k}^{k,t}\}$ for $k = 1$ to K , the algorithm reports the partition $B^t = \{B_1^t, B_2^t, \dots, B_{\tilde{c}}^t\}$ that satisfies

$$B^t = \left\{ \max_{B_x^{k,t}} \mathbf{E}_{l \neq k} \left[\max_y H(B_x^{k,t}, B_y^{l,t}) \right] \right\}$$

where $B_x^{k,t}$ is the set of nodes in the x th block of the k th partition at time t , c_k is the number of blocks in partition k , \tilde{c} is the median number of blocks across all K partitions of time t , and $\mathbf{E}_{k \neq l}[\cdot]$ is the expectation taken over all pairs of different iterations. Here we use similarity function $H(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$.

This resolution step can also be extended to “smooth” the output by incorporating $B_y^{l,t-1}$ and/or $B_y^{l,t+1}$ (as well as additional time segments in each direction) into the bounds of the expectation, emphasizing connections between clusters over time if an algorithm with memory is desired. This smoothing approach compares the possible partitions available for time step t to those from $t - 1$ and $t + 1$ in order to find blocks which are consistent across different time steps.

A slight overhead is incurred during the resolution of the ensemble of partitions. For a given time step, each partition k out of the K different partitions of the graph has c_k blocks, and all pairs of blocks from different partitions must be compared for the intersection of their member nodes. If using a hashed implementation, each intersection operation costs $\Theta(\min(n_x, n_y))$, where n_x and n_y are the size of blocks x and y . For every pair of partitions, all pairs of block must be compared, and each are an average size of $\frac{N}{c_k}$ or $\frac{N}{c_l}$. The cost of all comparisons is then

$$\Theta \left(c_k c_l K(K-1) \min \left(\frac{N}{c_k}, \frac{N}{c_l} \right) \right) = \Theta(N \cdot \min(c_k, c_l)).$$

Assuming that the number of blocks returned in any partition is less than $\log^2 N$, the computational complexity of the ensemble resolution step is equal to or less than the cost of clustering the graph once at any particular time segment on average. Each iteration of the graph clustering algorithm in the ensemble can be run in parallel, needing no communication across the iterations until the resolution step, making the cost of each time segment $\Theta(N \log^2 N)$. In the memoryless, unsmoothed version of the framework the individual time segments are also independent and can be run in parallel. The smoothed version requires communication between adjacent time segments after graph clustering but before the cluster resolution step, where the clouds of clusterings obtained by the K iterations are exchanged.

2.3. Visualization of Output

Finally, after producing a representative clustering at each time segment with smoothing between adjacent segments, we produce a visualization showing how the community structure changes over time. An example of the

visualization obtained from a real world network can be seen in Figure 7. In this community graph, the nodes represent communities with the node diameter indicating community size, and the communities are connected in sequential time segments if they share members with a thicker edge indicating more overlap between the two sets of members. This visualization allows for a greater understanding of the evolution of community structure than merely reporting on the number of communities at each time segment. Through it, community evolution can be observed, including events such as communities merging, splitting, forming out of background noise, and dissolving into background noise.

3. Experiments

To validate our method we performed a series of experiments on synthetic, semi-synthetic, and real-world datasets. We evaluated the clustering methods using both pairwise-precision and pairwise-recall, as well as number of blocks reported where applicable. The pairwise-precision measures the percentage of node-pairs which are correctly identified as being in the same cluster: (true positives)/(true positives + false positives). Pairwise-recall considers all pairs of nodes belonging to the same ground-truth cluster and measures the fraction of them which are in the same identified cluster: (true positives)/(true positives + false negatives). More information can be found in [14], [15].

3.1. Synthetic Experiments

The first experiment comparing the ensemble model to the baseline is performed on a synthetic dataset where edges emerge from a network stream over time and the clustering uses all edges observed up to the current time. The block model from which the edges are sampled is considered constant; observing more edges merely increases the sample size and considerations like model drift are ignored. The data consists of 500 node synthetic graphs: each of these has 8 ground truth blocks with a 1:5 external/internal edge ratio, and the number of edges observed was scaled from 1000 to 1900.

The results from this experiment are shown in Figure 3. Red lines represent the baseline MCMC algorithm while blue lines are the ensemble algorithm; smoothing was not used as it is designed for dynamic models not static ones. The dashed lines show the standard errors over ten runs of the algorithms. The left-hand panel reports the number of clusters found by both algorithms. Here, the black line shows the true number of clusters. The middle panel and right-hand panels report the pairwise-precision and pairwise-recall, respectively, as more edges are sampled. With fewer than approximately 1400 edges sampled neither method has good performance due to the number of blocks being under estimated. However, as more edges are sampled, the ensemble method has strictly superior pairwise-precision. The pairwise-recall performance appears about the same for both the baseline and the ensemble.

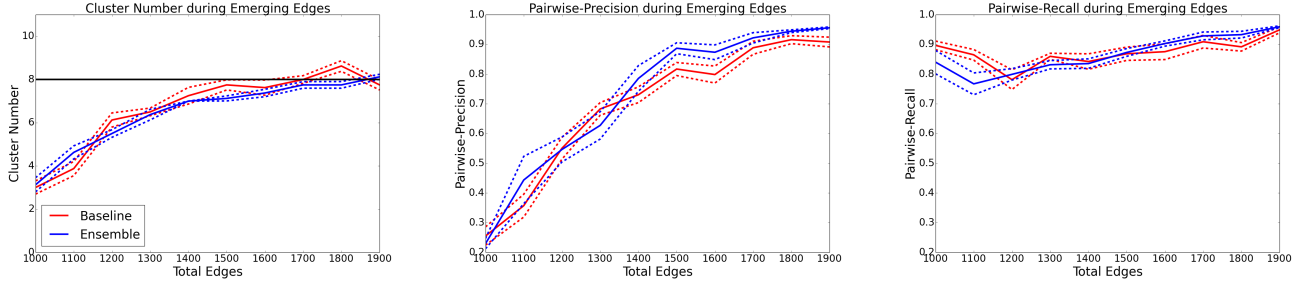


Figure 3: Performance on emerging edges experiment. Red lines represent the baseline algorithm and blue represent the ensemble approach, with dashed lines showing the standard error over 10 runs of the algorithms. The left-hand panel shows the number of blocks reported (compared to ground truth shown by the black line) as more edges are sampled. Pairwise-precision among nodes as more edges are sampled is shown in the middle panel. The right-hand panel shows the change in pairwise-recall among nodes as more edges are sampled.

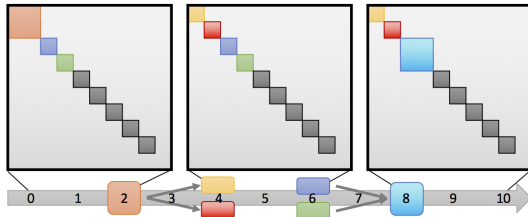


Figure 4: A high level view of the evolving community dynamics from the second experiment on synthetic data. The blocks are shown in the “spy plot” of the adjacency matrices associated with the underlying block model.

The second experiment is another synthetic dataset, but here the edges are drawn from a dynamic model where the block structure is changing over time. The model consists of a 100-node block which splits into two 50-node blocks between time segments 2 and 4; two 50-node blocks which merge into one 100-node block between time segments 6 and 8; five 50-node blocks which remain constant; and 50 nodes which connect to any node with equal probability, for a total of 500 nodes. 2000 edges were sampled for each time segment. A high level view of these dynamics is shown in Figure 4.

The results of this experiment are shown in Figure 5. Here, the red and blue lines represent the baseline and ensemble algorithms, respectively, as in Figure 3. The purple lines represent the ensemble with smoothing. The gray regions represent times when the block structure is in flux; the ground truth is set to be that of the blocks at the end of the transition, hence the drop in performance at the start of each region when the behavior observed has yet to fully align with the new ground truth.

The left-hand panel in Figure 5 reports the number of clusters found by the various algorithms used. The black line represents the number of clusters in the ground-truth clustering. The middle panel reports the pairwise-precision over time and the right-hand panel shows the pairwise-recall.

At the start of the split event, the precision drops as the ground truth has more clusters than the immediate graph

structure (until the cluster has fully split). As the clustering algorithm will misidentify nodes in the splitting cluster as being in the same community, the precision suffers. Likewise at the start of the merge event the number of communities is underestimated, placing members of the merging communities as separate which causes recall to drop.

For number of blocks found, the ensemble method performs slightly better than the baseline, especially in more quickly identifying a split. The smoothed ensemble matches the ground truth cluster number almost exactly. In terms of the pairwise-precision, the ensemble algorithm has strictly superior precision to the basic algorithm while the smoothed algorithm is better still. The same relationship holds for recall but the difference is not statistically significant at all times.

3.2. Semi-Synthetic Experiment

The third experiment is performed on a semi-synthetic dataset which was produced by inserting nodes and edges sampled from a block model into a real-world network, the “email-Eu-core-temporal” dataset from the SNAP database [16]. The synthetic blocks inserted consist of a 100-node block which splits into two 50-node blocks between time segments 2 and 4 and two 50-node blocks which merge into one 100-node block between time segments 6 and 8. This is similar to the dynamic structure in the previous experiment, except that now nodes from the synthetic blocks are randomly connected to nodes in the real-world email dataset, rather than nodes from other synthetic blocks. This more accurately measures the ability of the algorithms to identify dynamic community structure under conditions which more closely represent those found in real-world data.

The synthetic data has 960 edges per time segment, 800 of those being between pairs of synthetic nodes and 160 going to nodes in the real-world data; the real-world data consists of the first 10 weeks of the dataset which has 986 nodes and 38330 edges, split fairly evenly across those weeks. Pairwise-precision and pairwise-recall are reported for the synthetic nodes only.

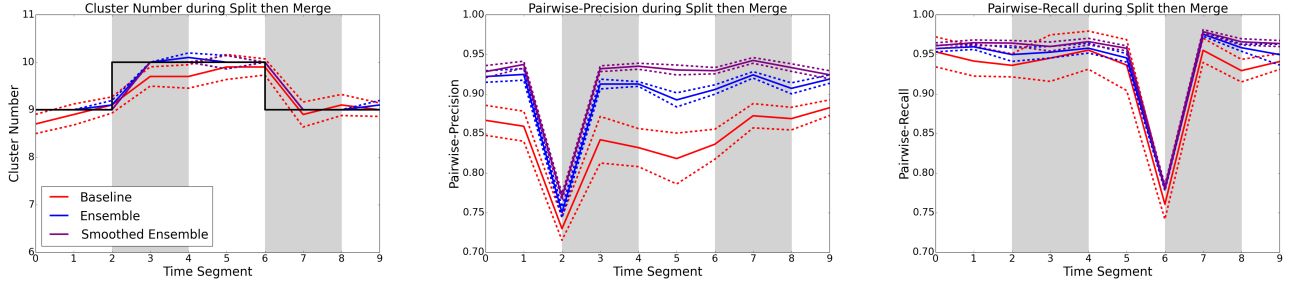


Figure 5: Performance on synthetic split and merge dynamic event. Red and blue lines represent the baseline algorithm and the ensemble, respectively, with dashed lines showing the standard error over 10 experiments, as in Figure 3. Purple lines, solid and dashed, show the same for the ensemble approach with smoothing between time segments. The grey regions represent time periods during which the underlying stochastic block model is changing. The left-hand panel shows the number of blocks identified in each time segment using the baseline and the ensemble as compared to the ground truth, shown by the black line. The middle panel shows the pairwise-precision for all three methods over time and the right-hand panel shows the pairwise-recall.

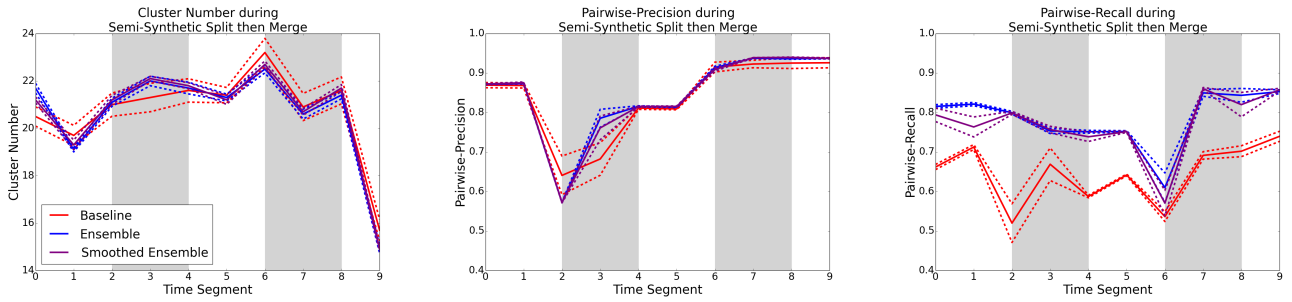


Figure 6: Performance on semi-synthetic split and merge dynamic event. Red, blue, and purple lines represent the baseline MCMC clustering algorithm, the ensemble approach, and the ensemble approach with smoothing, respectively, with the dashed lines showing the standard error over 10 runs. The grey regions represent time periods during which the underlying stochastic block model on the synthetic nodes is evolving. The left-hand panel shows the number of clusters found by each algorithm over time, the middle panel shows the pairwise-precision for the synthetic nodes, and the right-hand panel shows the pairwise-recall for the synthetic nodes.

The results of this experiment on semi-synthetic data are shown in Figure 6. As previously, the red, blue, and purple lines represent the baseline MCMC, ensemble, and ensemble with smoothing algorithms, respectively. The left-hand panel shows the number of clusters reported by all three algorithms. Clusters discovered in the real data are included in this count, which is the cause of the drop in cluster number in the final time segment: two large merges occurred in the real dataset at this time reducing the overall number of clusters. These merges are visible in the community structure shown in Figure 7.

The grey bars highlight the time segments in which the clusters of synthetic nodes are undergoing a split or merge. Unlike the previous experiment, the ground-truth number of clusters is not reported due to the fact that the ground truth communities are not known *a priori* for the “email-Eu-core-temporal” dataset. The middle and right-hand panels show the pairwise-precision and pairwise-recall over time. In this experiment the pairwise-precision does not seem significantly different across the different algorithms, however the recall for the baseline algorithm is much worse than either the ensemble or the smoothed ensemble.

3.3. Real-World Experiment

Finally, we ran the smoothed ensemble framework on the first 10 weeks of the EU email dataset. The dynamic community structure obtained is shown in Figure 7. The horizontal axis represents time (in weeks), with community detection taking place weekly. The points represent the communities found during each week, with the diameter of the points representing the size of the communities at a given time and the thickness of the lines between communities in different weeks representing the node set overlap between two communities. Many of the communities are constant, forming long horizontal lines with no splitting or merging behavior. Others undergo prominent splitting and/or merging behavior. There are also a number of communities which appear to “die” over the course of the ten weeks, with the community terminating without merging into another. Of particular note are the pair of large merge events occurring in the final time segment where multiple communities join to form large blocs.

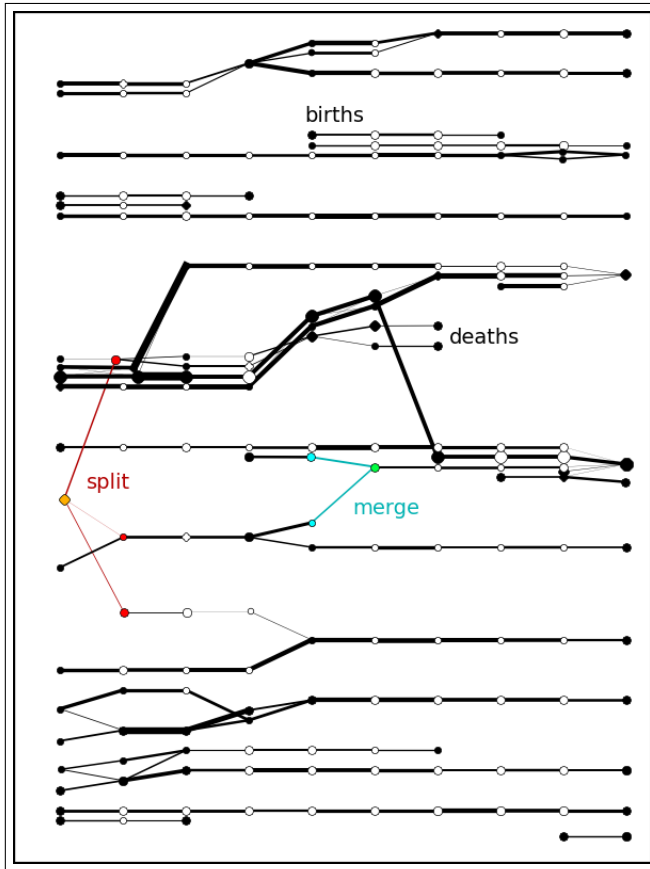


Figure 7: Timeline of the clustering behavior of SNAP European University dataset over the first 10 weeks of data using the ensemble algorithm with smoothing.

4. Conclusions

The ability to perform dynamic clustering on real-world networks and produce an interpretable visualization is a powerful tool for network analysis. The ensemble framework presented in this paper has the ability to produce high-level summations of the evolving community structure of dynamic networks and can be applied to large networks due to its efficient parallel implementation. We also presented a version of the ensemble with memory which allows for more stable representations of the community structure. We demonstrated using synthetic, semi-synthetic and real-world scenarios that the ensemble algorithm has superior precision and recall performance compared to the basic clustering algorithm, and we also presented the types of dynamic community representations that the ensemble algorithm can produce from real graphs.

Future directions for this framework include incorporating more information from previous time segments when performing clustering: for example, edges from graphs in previous segments could be included in the current graph using an exponential decay scheme. This would allow a more narrow resolution of the time segments without making the

graphs so sparse as to be useless. Other possible improvements exist for the visualization of the dynamic network: for example, the variability in the node membership of the representative clusters could be visualized to represent the level of “confidence” the algorithm has in a particular cluster assignment.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

- [1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” *Physics Reports*, vol. 424, pp. 175–308, 2006.
- [2] U. Brandes and T. Erlebach, eds., *Network Analysis: Methodological Foundations*, vol. 3418 of *Lecture Notes in Computer Science*. Springer, 2005.
- [3] M. E. J. Newman, *Networks: An Introduction*. Cambridge University Press, 2010.
- [4] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.
- [5] F. D. Malliaros and M. Vazirgiannis, “Clustering and community detection in directed networks: A survey,” *Physics Reports*, vol. 533, no. 4, pp. 95–142, 2013.
- [6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 10, p. P10008, 2008.
- [7] T. P. Peixoto, “Efficient monte carlo and greedy heuristic for the inference of stochastic block models,” *Physical Review E*, vol. 89, no. 1, p. 012804, 2014.
- [8] T. P. Peixoto, “Parsimonious module inference in large networks,” *Physical Review Letters*, vol. 110, no. 14, p. 148701, 2013.
- [9] T. P. Peixoto, “Entropy of stochastic blockmodel ensembles,” *Physical Review E*, vol. 85, no. 5, p. 056122, 2012.
- [10] D. Greene, D. Doyle, and P. Cunningham, “Tracking the evolution of communities in dynamic social networks,” in *International Conference on Advances in Social Networks Analysis and Mining*, 2010.
- [11] J. Sun, S. Papadimitriou, P. S. Yu, and C. Faloutsos, “Graphscope: parameter-free mining of large time-evolving graphs,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2007.
- [12] Y. Hong, C. Fu, Q. Huang, Z. Fang, J. Zeng, and L. Han, “Communities evolution analysis based on events in dynamic complex network,” in *IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, 2016.
- [13] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, “Netprobe: a fast and scalable system for fraud detection in online auction networks,” in *Proceedings of the 16th international conference on World Wide Web*, pp. 201–210, ACM, 2007.
- [14] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
- [15] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, p. 193218, 1985.
- [16] J. Leskovec and A. Krevl, “SNAP Datasets: Stanford large network dataset collection.” “<http://snap.stanford.edu/data>”, June 2014.