The Comparison of Graphical Algorithms in Network Science for the Diffusion of Information in Social Networks

Murthy, Pranav
College of Computing
Georgia Institute of Technology
Atlanta, Georgia
pmurthy32@gatech.edu

Stone, Nicholas
College of Computing
Georgia Institue of Technology
Atlanta, Georgia
nstone31@gatech.edu

Reddy, Saketh
College of Computing
Georgia Institute of Technology
Atlanta, Georgia
sreddy347@gatech.edu

***Abstract*—We address the existing literature to construct a distinction between the terminology of community and clique, community referring to a set of nodes within a social network which contain a higher probability of intersecting. This is representative of groups of people by affiliation, relation, or common interests. On the other hand, we define cliques as a subset of communities–a smaller community, if you will–but a community that is wholly connected; each node is clustered through intersection. To be a clique, there must be a node intersection. We assert that there exist noticeable trade-offs in terms of accuracy and stability of graphical models, such as: The high computational complexity of the Clique Percolation Method in combination with its high interpretability as a direct result of its integration of multi-clique-based data, the use of non-negative matrix factorization through the discovery of latent features and representation by parts, while accounting for lack of scalability and a standard objective function. With this project, we seek to document these tradeoffs by generating concrete visualizations of these adjacent variances.**

I. Introduction

The discipline of Network Science refers to the complex study of entities, or actors, across varied situational relationships such as biological, semantic, or social networks. The study represents actors through a series of nodes, and defines a certain connection between two nodes as a link, or edge. Inspired upon ideas such as graph theory and data mining, the overarching purpose of Network Science as a means to interpret a network of relationships is to understand how biological, social, and situational/environmental stimuli or predispositions can be used as a predictor of certain behaviors. In particular, the study of social networks has emerged as an increasingly more relevant topic given how the ever-evolving landscape of technology informs the introduction of mediums which transform how entities interact socially amongst peers. Consequently, the existing various methods used to assess social networks have been increasingly debated as to which method best optimizes despite its computational limitations to predict actor interactions.

Certain trade-offs in regards to interpreting more compounded data go hand-in-hand with higher computational complexity, indicative of a much steeper learning curve, but a more accurate and feasible construction for social interactions. One particular distinction in the realm of simulating social networks is how a certain method of simulation deals with the inclusion or disregard for cliques, a subset of a node community which is much more precise in monitoring distinct interactions between nodes, and one that is wholly connected (i.e. to be a clique, each node is clustered via an intersection across all other nodes in that community). The inclusion or exclusion for many of these tradeoffs primarily deals with computational complexity and the laborious or overly-elaborate hardships that occur when attempting to scale a simulated scenario to a more macro context.

For many researchers making use of Network Theory in their assessment of a given social network, we assert that the comparison of various graphical representations of social network intricacies will enable students and other professionals new to the field to quickly develop a stronger and more conceptual understanding of these critical techniques within Network Theory.

1) Gap: As outlined above, Modularity Maximization, Non-Negative Matrix Factorization, and the Clique Percolation Method are powerful graphical functions for extracting node relationship information from connected graphs, defining classifications such as communities and cliques that can enable powerful tailored insights into segments of the graph. Regardless, little to no research actually exists on the precise implications in which communities and cliques must be factored in when researching a given social network; there does not exist a strong method for students to effectively visually compare these two different graph clustering techniques and conceptually understand their differences and limitations of these different methods. This project seeks to resolve this resource gap through the creation of visualizations accompanied by minimal reproducible code that can be used to efficiently compare and contrast the outputs of these techniques.

Special thanks to Professor Gerandy Brito.

II. Background and Existing Literature

A. Computational Complexity

1) Inclusion of Clique Data Versus Generalized Community Data: In the context of social networks, a community describes some series of nodes that have a non-zero probability of intersecting. These nodes have some overlapping similarity amongst each other: This may be based in the context of varied relationship statuses, interests, or otherwise. These large trends across actors are documented to portray various trends of node interactions at a macro level, but not every node needs to have a link to each of the nodes in the community. So long as there is a distinct and denser internal connection between a series of nodes, a community can be defined. A clique, on the other hand, is a community which refers to some subset of vertices within a graph that are wholly connected, meaning that each node must have a link to every single other node in the included community. In the realm of Network Science, a clique of nodes would describe a certain sub-group of people within an entire social network. Cliques play a part in community interactions, and provide a much more complex and holistic delineation for how each node functions at a much more precise level.

The idea of k-cliques describes a fully connected group of cliques with k nodes in each clique. The term k-clique adjacency is defined if two k-cliques are adjacent based on sharing k - 1 vertices. A k-clique chain is a generated graph of k-cliques where each clique is represented next to another, with each edge representing the strength of similarity from one clique to the next. Two cliques in the same chain are known as k-clique connected with each other. A k-clique percolation cluster can be described as a graph where each node represents a certain k-clique, which allows for a deeper understanding of how the sub-groups interact with each other. The cluster must be composed of only cliques that are k-clique connected. Each edge represents relative strength concerning the overlap from one clique to the next.

B. Clique Percolation Method

The clique percolation method (CPM) is a unique approach applied to social networks, particularly to Erdős Rényi (ER) uncorrelated random graphs, that are based on the k-clique approach of a pre-defined number of sub-groups within a social network where all cliques are fully connected. In terms of social networks, CPM detects the existence of public communities through k-cliques where each node would represent an individual. Each edge generated by an ER graph has an independent probability $0 < p < 1$. Given an ER graph with N vertices and each k-clique containing $\begin{pmatrix} k \\ 2 \end{pmatrix}$ edges, the expected number of k-cliques would be $C = \begin{pmatrix} N \\ k \end{pmatrix} p^{\binom{k}{2}}$, given $\begin{pmatrix} N \\ k \end{pmatrix}$ being the total number of possible cliques that could be formed. New k-cliques can be formed from an initial i-clique, containing i nodes. The resulting probability would be $p^{\binom{k}{2}-\binom{i}{2}} = p^{\frac{(k-i)(k+i-1)}{2}}$. This can be used to determine the probability distribution of all new k-cliques formed from the i-clique [2].

This distribution is critical in the context of social dynamics of a graph. Because cliques within a network act as information repositories for each individual, the probabilities of each k-clique alters the structure of each cluster focal point. Moreover, there is a concept in sociology known as homophily, which describes the concept of people being more inclined to "seek out" those that are similar to them. This idea can also be applied to nodes and cliques, where the probability distribution determines how similar the cliques formed are from each other. If cliques are closely related in structure, size, and/or density, there is a higher chance that information will disseminate more effectively.

It is also important to note the distinction between homogeneous and heterogeneous networks, which ties back to homophily within a social graph. The probability distribution determines if information could spread through diverse or similar groups. If the k-cliques had some diversity within their organization, any intelligence that is spread will reach cliques with varied forms. Information will flow better within similar k-cliques, and a homogeneous network contains an even distributional structure of cliques, which could potentially lead to what is known as an "echo chamber." Echo chambers are environments where individuals are exposed only to the same data that coincide with their own beliefs. In a mathematical network of individual nodes, this may also unfold.

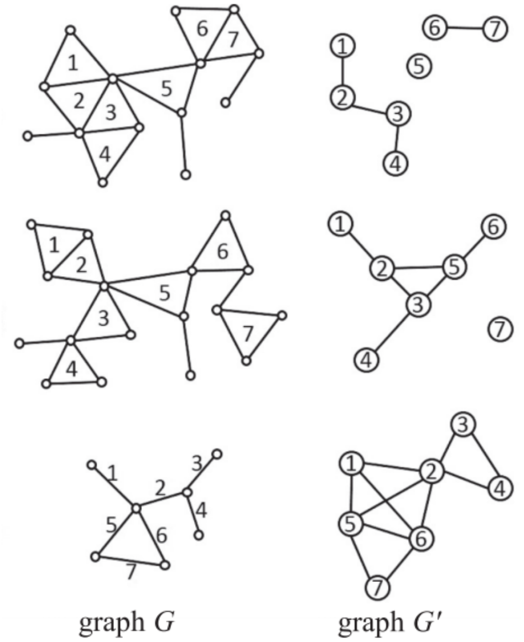

Fig. 1. k-cliques in a graph translated through CPM

Figure 1. [2] shows how the clique percolation of a graph G can be viewed as normal percolation in the corresponding graph G'. Each clique formed is represented as a node counter-

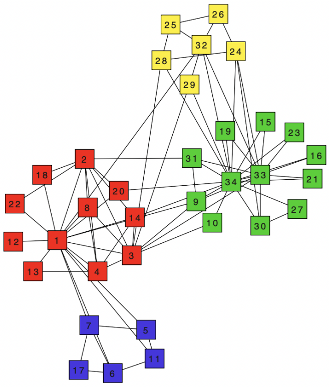

Fig. 2. Zachary's Karate Club, baseline observation of MM [1]

part, essentially simplifying the clique dynamic in the original graph.

C. Modularity Maximization

Modularity Maximization is a graphical function that discreetly identifies communities under the guise of meta-nodes, or collections of nodes that have collective functioning similarities, but zero intersection to other node communities. Each node is evaluated using the most necessary and relevant connections to the applied context of the research, with subsequent edges being disregarded, which does add a non-zero design error, but one that is essentially non-existent, and one that is often taken on to shear the added complexity that more edges create. In short, Modularity Maximization seeks "a partition of the graph into disjoint communities, i.e. into sets such that each node belongs to exactly one set" [1]. These partitions are often interchangeable with "clustering," and as such, one interpretation of observing clusters was determined by Newman and Girvan by maximizing the modularity of a given meta-node. Figure 2 indicates a baseline example that is commonly used to visualize Modularity Maximization. Commonly known as Zachary's Karate Club, the figure is representative of a group of karate students that were split apart into two separate karate studios across a year of observation (Agarwal and Kempe). Modularity Maximization clearly distinguishes the four communities of students: those who attended the red studio, those who left the red studio after the split (In blue), those who attended the green studio, and those who left the green studio (in yellow). A clear delineation between communities is identified, allowing for a scalable and clear means of interpreting the karate data.

D. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is another approach to tackling this problem. NMF describes the method where, given a non-negative matrix (pxr), referring to a matrix in which all elements are strictly greater than zero, two resulting factorized matrices are continually decomposed into smaller matrices (pxq and qxr) until each pair can be represented by the resulting mxn matrix. In the context of a graph made up of nodes and node clusters, it is possible for the method to convey patterns across each factorized matrix into another, implying node connections from one clique to the next. NMF may potentially allow for easier interpretation of clusters, and subsequently, the extraction of particular qualities for a subset of people. However, unlike CPM or modularity maximization, NMF is typically not considered in the field of network science. Despite there not being much specific research into its applications for graphs, NMF allows for a new perspective on how clustering within a network occurs.

Given a matrix $A \in \mathbb{R}^{m \times n}$ and an integer k such that $k < m, n$. The NMF method will decompose this matrix into $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{n \times k}$, so $A \approx WH^T$. To find the values of W and H, the difference between A and WH^T must be minimized, as long as both matrices are positive. It is important to note that the true minimum cannot be solved, so a local minimum must be determined. Thus, the following equation

$$\min_{W,H} f_k(W, H) \equiv \frac{1}{2}||A - WH^T||_F^2$$

must be satisfied [3]. This is typically done through gradient descent, where gradually going through this function will yield a minima for our result.

III. Method

We proceed using a method of modeling to observe how the various graphical functions as noted above take in similar parameters and interpret their community and, if applicable, subsequent clique relations.

A. Erdős Rényi Model

Within Network Science and graph theory as a whole, simulating networks of nodes requires a model that is capable of generating graphs such that the information within them (distance between nodes, interaction of cliques, etc.). The Erdős Rényi (ER) model allows for both generating and evolving random graphs over time. The ER model deals with uniform probability for a certain immutable number of edges between nodes, as well as a fixed number of individuals overall. That is, each edge has an equal likelihood of existing, where the existence of each edge does not affect another.

B. Implementing Clique Percolation Method

When simulating CPM, a method to find the cliques for a given graph must first be implemented. This was done through initializing a list of sets that each contain some number of nodes that is at least the value of some parameter k. Then, this was passed into the clique percolation function, which determines the pattern of cliques. The following code describes how this was achieved. The function operates where, given a graph and some value k, a function is called to find all the cliques for the graph. Then, a new empty set is initialized, where every clique is compared with another in the graph. If both cliques are unique and their intersection contains at least $k - 1$ nodes, then the union of both cliques are added to the new set. After parsing through all cliques, the final set is returned.

```
function CLIQUEPERCOLATION(graph, $k$)
    $cliques \leftarrow$ FindCliques(graph, $k$)
    $cliqueSet \leftarrow$ EmptySet()
    for $clique1$ in $cliques$ do
        for $clique2$ in $cliques$ do
            if $clique1 \neq clique2$ and $|\text{clique1} \cap \text{clique2}| \geq (k-1)$
then
                $cliqueSet \leftarrow cliqueSet \cup (\text{clique1} \cup \text{clique2})$
            end if
        end for
    end for
    return List($cliqueSet$)
end function=0
```

C. Implementing Modular Maximization

Building the MM algorithm involved leveraging code from the NetworkX module in Python. Essentially, the algorithm first initializes each node as its own "community." Then, for each neighboring community of a given node, determine the potential change in modularity if that node moved to that specific community. Then, move the node to the community for which the highest modularity change occurred. Repeating for each node, the final communities are then returned.

D. Implementing Non-negative Matrix Factorization

Constructing an NMF algorithm for the randomly generated graph required some way to convert an ER graph into a non-negative matrix. This was done through what is known as an adjacency matrix, where given two values in a matrix representing two nodes, the value between them is either 0 or 1 for if there is a connection between them. After applying this matrix to the NMF function from the sklearn module in Python, we can fit the resulting matrices back to a network graph.

IV. RESULTS

The ER graph has been been initialized, as seen in Fig. 2. This graph in particular effectively demonstrates diversity in its structure, with the number of connections per node being very heterogeneous. There are nodes ranging from outlier individuals with no connection, nodes being connected only through one relation, and nodes being connected to four others.

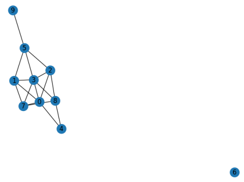

Fig. 3. Randomly initialized ER Graph

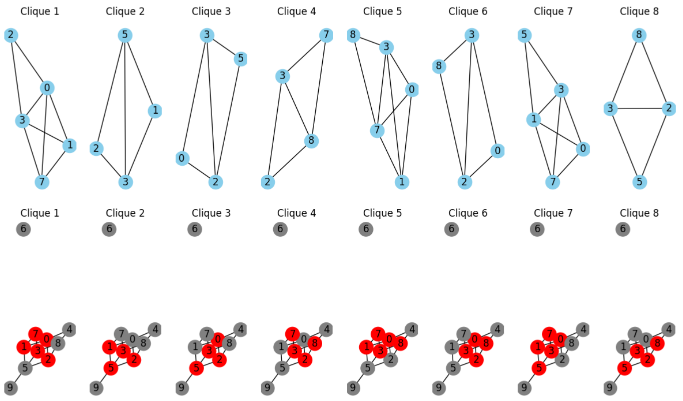

Fig. 4. Communities from Fig. 2 generated through CPM

A. Clique Percolation Method

Based on the results shown in Fig. 3, there are 8 cliques that were initialized. Each initialized clique shown indicates a sub-relationship based on a given similarity observed within the initialized scenario, detailing the varied relationships that occur within a shared community. Based on the structure of each clique, it is fair to say the network is heterogeneous, and so any information that may be spread would be exposed to diverse groups.

B. Modularity Maximization

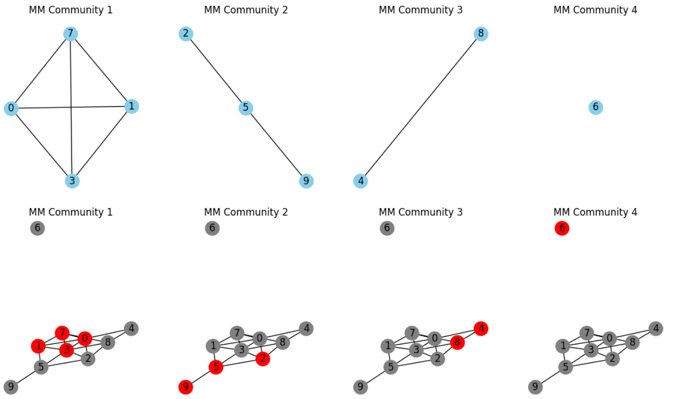

Fig. 5. Communities

As observed within the compilation of the ER graph through Modularity Maximization, rather than denoting the tightly-knit micro relationships that are observed by cliques, the broader communities are observed, allowing for a much more digestible and readable observation of the sects of a given social network. Notice how CPM failed to account for node 6, a completely isolated node from the rest of the social network. Modularity Maximization distinctly identifies the overarching trends of communities by forgoing the more precise node interactions that often boil down to expected behavior. In the context of a scalable social networking project, this may be why Modularity Maximization as a medium of analyzing is preferable given how each of the clique interactions in CPM are directly showcased in the ER Graph, but Modularity Maximization shears the expected data off of the edges to showcase the macro node interactions at play.

C. Non-Negative Matrix Factorization

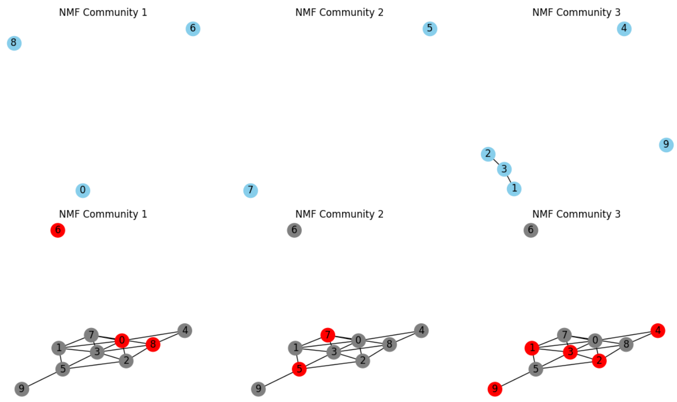

Fig. 6. Communities from Fig. 2 generated through NMF

The purpose of the implementation of Non-Negative Matrix Factorization is as an unorthodox approach to observing the relationships of a given social network. That being said, the visuals as observed in Figure 5 for NMF do not particularly expose relevant connections between nodes. In fact, it only detailed the non-relationships between series of nodes, besides its third iteration, where it cause a community between nodes 2, 3, 1. NMF did not particularly expose any relevant connections between nodes and proved to be a particularly ineffective and unreliable form of catching the intersections between nodes through community and clique relations.

V. Extensions and Applications

A. Limitations through Generalizations

Some limitations do need to be addressed regarding the simulation of social networks. Despite networks of people behaving in unpredictable ways, assumptions must be made regarding groupthink and defining discrete probabilities of actions that a node can make which reduce the computational complexity of a given graphical function and make its assessment a feasible option of simulating social-adjacent behavior. For example, there will most likely be uniform probability across the different person nodes in a simulated social network. These probabilities will define not only how groups of interconnected people would function graphically, but individual nodes as well, which may not be very similar to how social networks operate in the real world, especially regarding the dissemination of data.

The Clique Percolation Method (CPM), for example, assumes that people are confined to specific cliques among a larger group, where one is connected to every other person within a group. This is generally untrue for groups and subgroups, specifically those that become larger over time, as a person can only be social with a certain number of people. CPM also required a fixed size for each community, and these overlapping simulations are far more predictable this way, rather than more nuanced groups that grow through functions over time. CPM also states that an individual node has the same connectivity weight for every person they know, whereas in reality, members of groups know some far more than others.

Modularity maximization is also a method of analyzing social communities, but it too comes with assumptions. Unlike CPM, this model does not consider overlapping communities in its calculation, meaning each individual belongs to only one specific subgroup. Further, it requires a predetermined number of communities set, rather than communities that spontaneously form much like real social networks. Similar to CPM, modularity maximization has some form of homogeneity among the connections between people, assuming that all connections have the same weight. Both CPM and modularity maximization assume the network structure remains constant as time progresses. The model also considers modularity, which is to what extent a network can be separated into communities, as the sole metric for the analysis of an entire network of people.

Non-negative matrix factorization (NMF) is a more broad method of finding patterns in a social community, but it is regardless a powerful tool for data analysis. However, one key generalization that NMF considers is that all input and output values are non-negative. This may prove to be a problem in a network of people where certain data can be from a range of positive to negative, such as sentiment analysis, and the perceived/objective trustworthiness of different people in certain groups. As stated previously, the growth of a social group over time may not always follow a linear prediction, but linearity must be a given for NMF, despite real world communities behaving differently.

B. Extensions and Applications

Despite the generalizations that these mathematical models make to analyze and predict social networks, these methods can still be very powerful when simulating the real world systems of people. While individuals act very unpredictable by themselves, a group setting can be simulated more accurately through probability of networks. The graphical models mentioned have the potential of determining important information on social structures.

One specific application that can be studied regarding diffusion across a network model is the spread of misinformation. Misinformation are false claims that are spread across the internet which has the purpose of deceiving the public to believe incorrect information. In the past few years with the advent of instant communication with anyone online, the spread of misinformation has had a wide range of negative consequences on real world events.

Mathematical models can help to analyze how false information spreads and potentially how to counteract it. Doing so could help with a more informed community and fewer negative implications both for people within and outside of certain networks. The Clique Percolation Method, for example, can help to identify communities that specifically spread this misinformation. Detecting the clique from which the information originated from allows for the model to find specific individuals within the subgroup that are actively diffusing the data. The modularity maximization can allow for determining the sentiments of individuals and groups as a whole, which may be a signal of misinformation being spread in certain directions. Non-negative matrix factorization can also help analyze user behavior by modeling content engagement to detect patterns, especially those that are engaging in misinformation, among groups.

Another application of social graphical modeling is detecting certain anomalies among a network, such as fraudulent events. Identity theft, bot accounts, and unusual behavior in general is necessary to detect when dealing with large scale networks of people, especially online through instant diffusion of information.

The mathematical approaches can help through a variety of methods for fraudulent analysis, such as influence mapping, which determines certain people within a social network who are more likely to behave in unusual ways or spread negative content. Models such as NMF can also help through cross-network analysis, which evaluates the data being spread through platforms, seeing which nodes behave differently to detect people who use multiple accounts online or automated scripts pretending to be people.

C. Future Work

While the models described have great potential for the analysis of social and communal networks, there are various more models that can potentially be researched further to provide insights and different perspectives for evaluation. Graph theory metrics is a great example of this; methods such as degree centrality, which measure the connections in a network per node to indicate prominence, or eigenvector centrality, which views the connectivity of a node and its neighbors' centralities, could find emerging data within a network. Network detection algorithms can be used as well.

The Louvain Method, which is closely tied to modularity maximization, can help to classify a total group of people into clusters not manually through cliques, but through recognizable patterns. The Girvan-Newman algorithm also has great potential in finding communities in a similar fashion by gradually removing any connections in the centrality of a group. Network embedding models can also help with applying a number of dimensions to nodes to better determine the structure of a network through various outlooks, and influence diffusion models can help with finding the spread of influence through a network, which is similar to the graph methods covered previously.

Acknowledgment

Murthy, Stone, Reddy thank Professor Gerandy Brito for his phenomenal work and guidance.

References

[1] G. Agarwal and D. Kempe, "Modularity-maximizing graph communities via Mathematical Programming," The European Physical Journal B, vol. 66, no. 3, pp. 409–418, 2008. doi:10.1140/epjb/e2008-00425-1

[2] M. Li, Y. Deng, and B.-H. Wang, "Clique percolation in random graphs," Physical Review E, vol. 92, no. 4, 2015. doi:10.1103/physreve.92.042116

[3] J. Kim, H. Park, "Sparse Nonnegative Matrix Factorization for Clustering," Georgia Institute of Technology, 2008.