COMMUNITY DETECTION IN LARGE NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

June Andrews

May 2012

COMMUNITY DETECTION IN LARGE NETWORKS

June Andrews, Ph.D.

Cornell University 2012

Networks are large and demand attention at being understood for these reasons.

With the impossibility of understanding a network a node at a time and the incompleteness of data, we seek to find clumps of data that exhibit cohesion. We call these communities.

With community detection we hope to better our understanding of large networks. This thesis makes advances towards understanding existing methods, introduces a greedy algorithm within current community detection and steps outside towards the creation of parallel community detection method.

Your abstract goes here. Make sure it sits inside the brackets. If not, your biosketch page may not be roman numeral iii, as required by the graduate school.

BIOGRAPHICAL SKETCH

June Andrews was born in San Diego, 1985. She attended University of California, Berkeley for her undergraduate degree in Electrical Engineering and Computer Science, with a minor in Applied Mathematics. She is now completeing her doctoral degree in Applied Mathematics at Cornell University.

Here’s to you Da.

”There are more things in heaven and hell, Horatio, than are dreamt of in your philosophy.” - the Bard.

ACKNOWLEDGEMENTS

It goes without saying, these people have been inspiring forces of nature to work with:

- Len Kulbacki
- Coach Wilson
- James Sethian
- Patricia Kovatch
- John Hopcroft
- Steve Strogatz
- Jon Kleinberg

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 History of Community Detection

Community detection has been subconsiously performed by scientists for decades. Marine biology studied the interaction between pods of dolphins [] with the idea that certain dolphins belonged to specific pods. The study of ... finds which species are similar and searches for a common ancestors to explain why groups of species are similar []. Blank studied the lunch time interactions between members of a karate club and found two groups emerged []. For these cases nodes are dolphins, species, and people, edges are interactions and similarities. And, just like that we are in the domain of network theory. For each of the examples, the graph sizes were small and conclusions could be drawn by intuitively assuming the communities. Once the communities were determined, larger view results could be obtained. That pods of dolphins interact in such and such a way, even though individual dolphins may vary from the overall conclusion. That a common ancestor to a set of species must exist, eventhough individual specie comparisons may appear more random. That a club already had a division, even though the resultant split was surprising.

The power of knowing the communities within a network is to draw conclusions on a much larger scale. Hence the study of communities commenses. The goal is that any application, that can draw a network representing their trade, can be decomposed into communities. If the decomposition can be done on the network, the human time to by hand understand all interactions and hypothosize what the communities are can be saved. Communities found, must be such that conclusions based on aggregate interactions between communities are meaningful. It is argueable that the kickoff to mathematically studying the detection of communities began with Girvan and Newman in [**?**].

From these origins community detection began, looking for how to partition a network into communities. Several avenues of research have been developed. We now briefly survey the major methods developed for finding communities.

The underlying assumption of these approaches, is that communities should have a dense set of connections within themselves, and as sparse as possible set of connections leaving the community.

JTODO: figure showing what is going on. Use a toy group, like karate or dolphins or football.

Notice that an intuitive grouping has a high number of *internal edges*, or edges between two nodes within the same community. Additionally, a community has a low number of *external edges*, or edges between two nodes in different communities.

1.2 Betweenness and Centrality measures

1.3 Metric Based Approaches

The question was, given a partition of the network, how good is that partition? Of course, the more distinct the partition, the more useful the found communities would be in coming to global conclusions. To measure the quality of a partition, several metrics exist. For each of these metrics then, finding the best partitioning is a matter of optimizing the metric.

The first use of a metric was in circuit board partitioning []. Since then modularity has become the overwhelmingly popular metric with other attempts as well.

1.3.1 Conductance

JTODO find a concrete paper on conductance

Conductance is a measure of the value of a cut within the graph. The lower the value of the cut, the more likely that one side of the cut is indeed a community []. In particular, for a given cut between C and $\bar{C}$, the conductance value of the cut is the ratio of external edges of C to all edges that have an end point in C. If C and $\bar{C}$ are disconnected, the value of the cut is then 0, the minimal value of conductance, the most likely senario for predicting the existance of a community C.

1.3.2 Modularity

Here is modularity.

1.3.3 Variations of Modularity

1.3.4 Other Approaches

1.4 Alpha Beta Clustering

1.5 Spectral Approaches

1.6 Approaches to Large Networks

1.7 Other Approaches

- Betweeness
- Swapping of Kernighan-Lin
- *k*-Clique Percolation
- Hierarchical
- Belief Propogation
- Principle Component Analysis

1.8 Desired Improvements

The most widely accepted community detection specific method is modularity maximization.

As data gets larger and more diverse, our understanding of what a community is, has changed. Additionally, partitions are no longer as meaningful.

Paragraph on what we desire to understand about larger networks

We deliver on the these results:

- The theoretical ability to compare existing community detection methods.
- A way to handle overlapping communities.
- A parallel algorithm with near perfect scalability to analyze arbitrarily large networks.
- An understanding of what network structure is.

For another time:

- way to algorithmically understand communities, once found

1.9 Notation

This thesis will use consistent notation and assumptions. Here, we provide a reference for all variables in table **??** and assumptions. There exist several similies in community detection, for clarity we mention them now. A network is also a graph. A node is also a vertex, person, paper, or any type of object within the network. An edge is also an interaction between two people, a citation between two papers, or a connection between any two objects within the network.

The assumptions we make are:

Table 1.1: Notation

Variable Name	Description	Constraints
V	Set of all nodes within the network	$\{u \mid u \in$ the network $\}$
u and v	Nodes	$u, v \in V$
$w(u, v)$	Edge Weight Function	$w : VxV \rightarrow \mathbb{R}_{[0,1]}$
G	Network or Graph	$G(V, E)$
C	Community	$C \subset V$
k	Fraction of nodes within C	$k = \frac{\|C\|}{\|V\|}$
$\|C\|$	Size of C	$\|C\| = k\|V\|$
S	Set of Communities	$S = \{C_1, C_2, \ldots, C_n\}$

Table 1.2: Introduced Functions

Function	Description
$I(C)$	Internal Density of a single Community, C, definition **??**
$E(C)$	External Density of a single Community, C, definition **??**
$I(S)$	Internal Density of a set of Communities, S, definition **??**
$E(S)$	External Density of a set of Communities, S, definition **??**
CONCISENESS(S)	Conciseness of a set of Communities, S, definition **??**

- *Self-Loops.* We presume there are no self loops in the networks. As a node will always be in the same community as itself, self-loops provide redundant information. Accordingly, $w(u, u) = 0$, for all $u \in V$.
- *Edges* We presume that all edges exist and are weighted between 0 and 1. The edge weight function is $w : VxV \rightarrow \mathbb{R}_{[0,1]}$ Unweighted graphs can

easily be adapted into this notation.

Definition 1 (Internal Edges) *Internal edges are edges between members of the same community C.*

Definition 2 (External Edges) *External edges are edges between a member of community C and a non-member.*

CHAPTER 2

A FRAMEWORK FOR THE COMPARING METRIC BASED DETECTION METHODS

2.1 Previous Comparisons

Given the variety of community detection methods, researchers have tried to compare them. In ..... they use conductance as a measure of the strength of communities produced by each method. In...

All comparisons have been of an experimental nature. A metric and data set are chosen and algorithms are compared. This methodology has lead to several comparison papers [] [] [], each of which offer general conclusions, but not theoretical results.

2.2 Individual Community Based Metrics

We now explore metrics that evaluate the strength of a single community. There are two uses of such metrics. The first is to recursively partition a network to find communities. This is done by finding the community, C, that maximizes the metric, partitioning the graph into C and $V - C$ and recursively partitioning the two subsets of the graph with the metric. The second, is to provide an intermediate way of evaluating the strengths of communities returned by more complex detection techniques. The later use is more common for these metrics, and for this purpose, CONDUCTANCE is the most popular intermediatory metric []. However, the benefit of using conductance in such a way has been unclear and is used more as an implicitly mediatory metric to compare communities from different detection methods than as an understood evaluation of communities.

In the following section, we show that the use of any of the existing metrics to find heirarchical partitions of a network will not result in conventionally accepted strong communities, confirming []. We also show that the use of these one dimensional metrics to evaluate communities, hides revealing information about the communities.

Our framework is to choose parameters that pertain to the desired characteristics of a community, parameterize the existing metrics, and analyze how change in the values of the parameters affect the metrics. By choosing parameters to understand the metrics, we can avoid the network dependent analysis of previous work [] and draw conclusions for all networks.

2.2.1 Internal Density and External Sparsity

It is accepted that the most distinct community is a clique, disconnected from the rest of the graph. There are two characteristics of this ideal community. The first is that it has high connectivity between the nodes of the community. We deem this property, *internal density*. The second characteristic is that the community has low connectivity to the rest of the graph. We deem this property *external density*. Formal definitions follow.

Definition 3 (Internal Density) *Internal density is the ratio between number of edges that exist between members of the community, internal edges, to the number of all* *possible edges that could exist within the community. Hence, $I(C) : C \rightarrow \mathbb{R}_{[0,1]}$, where*

$$I(C) = \frac{\sum_{u \in C} \sum_{v \in C} w(u, v)}{|C|(|C| - 1)}. \tag{2.1}$$

For a community C that has no edges between its members, the *internal density* will be minimized with, $I(C) = 0$. For a community C that is a clique, *internal density* will be maximized with $I(C) = 1$. Intuitively, the closer a community, C is to an *internal density* value of 1, the more sure we are the nodes have a meaningful connection that be used to observe aggregate behavior. If the graph is indeed a weighted graph, where w has values between 0 and 1, the same intuitions apply and *internal density* reflects how close a community is to being a maximally weighted clique.

Definition 4 (External Sparsity) *External sparsity is the fraction of edges that exist between a member of the community and a non-member of the community, external edges, to all possible edges that could exist leaving the community:*

$$E(C) = \frac{\sum_{u \in C, v \notin C} w(u, v)}{|C|(|V| - |C|)}. \tag{2.2}$$

For a community C that has all possible *external edges*, external density will be maximized at $E(C) = 1$. For a community C, disconnected from the rest of the graph, external density will be minimized at $E(C) = 0$. Intuitively, the closer a community is to having a value of $E(C) = 0$, the more complete the community is guaranteed to be.

There are other representations of $I(C)$ and $E(C)$ that vary the how the $|C|$ and $|V|$ terms are used. The analysis and conclusions that follow are not sensitive varying such variations. With our parameterization, we have that all communities can be mapped to a point $(I(C), E(C))$ that is in the square $\mathbb{R}_{[0,1]} x \mathbb{R}_{[0,1]}$, this is an easy to visualize space. We can now dictate what the ideal, or strongest, community is mathematically.

Definition 5 (Ideal Single Community) *A community, C, is ideal if it is an isolated clique, specifically has the following properties:*

$$\begin{aligned} I(C) &= 1 \\ E(C) &= 0. \end{aligned}$$

Metrics provide a one dimensional analysis of communities. Our proposal is to look at these two dimensions of communities explicitly, understand how metrics use these characteristics in their evaluation of a community, and from there draw our conclusions about metrics.

2.2.2 Study of Relevant Metrics

Given that we can map a community, C, to the point $(I(C), E(C))$, we now analyze how different metric based detection methods operate in the I, E plane. We cover six metrics that evaluate a single community. We use one approximation to simplify the equations, $|C| \approx |C| - 1$. This approximation has a larger impact on smaller communities, but most communities of interest are large enough to accomodate the approximation. Additionally, we introduce variable, k, representing the fraction of the nodes within community, C, such that $|C| = k|V|$

- CONDUCTANCE is the probability that a step in a random walk will leave the community [].

$$\textsc{Conductance}(C) = \frac{(1-k)E(C)}{kI(C) + (1-k)E(C)} \tag{2.3}$$

- CUT RATIO is the fraction of existing to possible edges leaving the community [].

$$\textsc{Cut Ratio} = k(1-k)|V|^2 E(C) \tag{2.4}$$

- EDGES CUT is the number of edges connecting the community to the rest of the graph [].

$$\textsc{Edges Cut} = E(C) \tag{2.5}$$

- EXPANSION the average number of edges leaving the community per node [].

$$\textsc{Expansion} = (1-k)|V|E(C) \tag{2.6}$$

- INTERNAL DENSITY as a metric, previously existed before our definition of $I(C)$, []. However, we stick to our definition of $I(C)$ for intutive reasoning and note in previous work internal density represents the mirror image of our definition.

$$\textsc{Internal Density} = 1 - I(C) \tag{2.7}$$

- VOLUME is the total degree of nodes within the community [].

$$\textsc{Volume} = |C|^2 I(C) + (1-k)|C||V|E(C) \tag{2.8}$$

Hence, we can put all previously described metrics in terms of $I(C)$, $E(C)$, $|C|$, $|V|$, and k. With a common parameterization of the metrics, we can already draw some inferences. All metrics, besides VOLUME are a function of either

$I(C)$ or $E(C)$, but not both. A metric that considers only $I(C)$ will be optimized by any clique. Which is a very restrictive definition of a community and finding all communities in the graph under such a definition is equivalent to finding all the cliques in a graph, a NP-hard problem. A metric that considers only $E(C)$ will be optimized by any disconnected component of the graph, including a community that includes the entire graph. While, it is possible to find all disconnected components in linear time, it also provides no information about reasonable datasets.

While these metrics are simple and easily understood by their parameterization, not all metrics have a closed form parameterization. We can push a step further and obtain a general methodology for gaining understanding of what communities certain metrics prefer. The methodology will be to use a simple greedy algorithm to incrementally improve a community according to a metric. The progress of the algorithm is then tracked within the (I, E) plane. While these paths may even seem random. We can visualize how a metric categorizes communities in the (I, E) plane with level sets. With the visualized categories in the (I, E) plane understanding the metric becomes a visual explanation.

The simple greedy algorithm is an expansion of the community process, greedily engulfing the node that improves a given METRIC the most. The algorithm is the *Greedy Single Community Metric Optimization* Algorithm: **??**.

Some metrics require minimization rather than maximization, this algorithm is easily adapted accordingly. For each metric we start the algorithm with a simple subset of two connected nodes and track the communities that result in improvements, according to a given metric. As the algorithm expands the community, the metric determines which node is optimal at each step. We record

Algorithm 1: Greedy Single Community Metric Optimization

```
Input: $C$, $G = (V, E)$, and Metric
  $inc = 1$
  while $inc \geq 0$ and $C \neq V$ do
    Let $u \in V$ maximize Metric($C \cup u$).
    $inc \leftarrow$ Metric($C \cup u$)$-$ Metric($C$)
    $C \leftarrow C \cup u$
  end while
  return $S$
```

this better community and for each metric, M, get a set of $C_1 \subset C_2 \subset \ldots \subset C_n$, such that the metric is increasing $M(C_i) < M(C_{i+1})$. While this is not an encompassing algorithm as it can only find local maximums, it will reveal a metrics biasenesses toward communities of certain $I(C)$, $E(C)$, values.

Let us look at how $E(C)$ based metrics categorize the (I, E) plane.

2.3 Set of Communities Based Metrics

Our original parameterization of internal and external density can not be directly applied to a set of communities, $S = \{C_1, C_2, ...C_n\}$. Though if we follow the same logic we will arrive at a similar parameterization. An ideal set of communities are cliques such that every edge is within some community. In addition, the community description of the network should not have an exponential number of communities, but rather some concise set of communities. Hence an ideal set of communities has three parameters. Internal density is a representation of how close the set of communities is to being a set of cliques. External density is a representation of how close the set of communities are to covering all edges in the graph. Size of the set of communities is a representation of how concise the set of communities are. We pick the following representations, but note that different representations do not yield different conclusions in the following sections.

2.3.1 Internal Density, External Sparsity, and Conciseness

Definition 6 (Internal Density of a Set of Communities)

$$I(S) = \sum_{C \in S} \frac{\sum_{u \in C} \sum_{v \in C} w(u, v)}{\sum_{C \in S} |C|(|C|1)} \tag{2.9}$$

Definition 7 (External Density of a Set of Communities)

$$E(S) = \frac{\sum_{u,v \in \textsc{Ext\_Edges}} w(u, v)}{\sum_{u,v \in V} w(u, v)} \tag{2.10}$$

Definition 8 (Conciseness of a Set of Communities)

$$\textsc{Conciseness}(S) = |S| \tag{2.11}$$

To date we do not know of any previous metrics that have a closed form parameterization in terms of $I(S)$, $E(S)$, and $|S|$. We release a linear metric pertaining to these parameters in a later section.

Definition 9 (Ideal Set of Communities) *A set of communities, S, is ideal if it is a set of cliques that cover the graph in the fewest necessary communities:*

$$\begin{aligned} I(S) &= 1 \\ E(S) &= 0 \\ |S| &= 1. \end{aligned}$$

2.3.2 Study of Related Metrics

CHAPTER 3

A NEW METRIC: LINEARITY

3.1 Single Community Detection

3.1.1 Results

3.2 Multiple Community Detection

3.2.1 Results

CHAPTER 4

PARALLEL COMMUNITY DETECTION

4.1 Introduction of Properties and Statistical Significance

4.2 Algorithm

4.2.1 Seeds

4.2.2 Expansion

4.3 Probability of Correctness

4.4 Performance

CHAPTER 5

CASE STUDIES OF NETWORKS

5.1 Amazon Product Network

5.2 Collaboration Networks

5.2.1 Astrophysics

5.2.2 Condensed Matter

5.2.3 High Energy Physics

5.2.4 General Relativity

5.3 Enron Email Network

5.4 Epinions Social Network

5.5 Gnutella P2P Network

5.6 Physics Citation Network

5.7 Web Graphs

5.7.1 Berkeley Webpage

5.7.2 Google

CHAPTER 6

EVOLUTION OF COMMUNITIES

CHAPTER 7

CONCLUSIONS

Above we have provided an indepth look at the details. Here we provide the summation of our results.

Finding communities is always a tradeoff. In metric based approaches between internal density and external sparsity. In significance based approaches the tradeoff is between specificity and sensitivity.

The number of communities a node belongs to follows a power law distribution.

Communities in citation networks evolve from a unioning of previous topics. However, not all papers that union topics produce successful communities.

[?] [?]

APPENDIX A

CHAPTER 1 OF APPENDIX

Appendix chapter 1 text goes here

BIBLIOGRAPHY