UNIVERSITY OF THE PHILIPPINES

Bachelor of Science in Computer Science

Aliya Ahlanna C. Miranda

W-CPMd: A Fast Clique Percolation Method for Directed Networks

Adviser:
Henry Adorna, Ph.D.
Department of Computer Science
University of the Philippines

May 2020

Thesis Classification:
F
This Undergraduate Research Paper is available to the public.

Aliya Ahlanna C. Miranda
Date of Submission: May 2020

Department of Computer Science
College of Engineering
University of the Philippines
Diliman, Quezon City

ENDORSEMENT

This undergraduate research paper hereto attached, entitled **W-CPMd: A Fast Clique Percolation Method for Directed Networks**, prepared and submitted by **Aliya Ahlanna C. Miranda**, in partial fulfillment of the requirement for the degree of **Bachelor of Science in Computer Science**, is hereby accepted.

HENRY ADORNA, PH.D.
Adviser

This undergraduate research paper is hereby officially accepted and approved as partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science**.

JAN MICHAEL YAP, PH.D.
Chair
Department of Computer Science

I hereby declare that I have created this work completely on I own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Aliya Ahlanna C. Miranda
Diliman, Quezon City
May 2020

Abstract

W-CPMd:
A Fast Clique Percolation Method for Directed Networks

Aliya Ahlanna C. Miranda
University of the Philippines, 2020

Adviser:
Henry Adorna, Ph.D.

Community detection algorithms are necessary for finding the relationships within a network. Real-time networks contains large amounts of data with a defined structure. W-CPM is a good algorithm for getting overlapping communities in large-scale networks due to its fast computational process but it is limited to undirected networks.

In this study, an algorithm that uses similar concepts to W-CPM but applicable to directed networks is proposed, called W-CPMd. If successful, this will help in distinguishing different communities in a vast variety of networks.

Table of Contents

Acknowledgments

To be written.

List of Tables

List of Figures

Chapter 1

Introduction

Community detection is the process of getting groups of interacting vertices, known as clusters or communities[4]. These clusters share common properties and specific roles in a network. By knowing these similarities, services can be improved like efficient storage data and navigation of queries[3]. It is also good for determining the structure of networks, like hierarchies. It is useful for getting target audiences or market in social networks and e-commerce businesses[4].

There are different types of communities. Disjoint communities are clusters that does not have common vertices, while overlapping communities have nodes sharing with multiple communities[4]. There are also directed and undirected networks, where in directed, relationships between vertices are not reciprocal, while it does in undirected. Creating algorithms for directed networks are a difficult task since it is characterized by asymmetrical matrices, thus its analysis is more complex[3].

In real networks, vertices commonly belong to more than one group, and highly organized with a hierarchical structure[3]. Clique percolation method, or CPM, is created by the concept that every vertex is connected to other vertex and forming a clique[4]. It is focused on undirected graphs by finding its k-cliques. This is extended for directed graphs in CPMd by formulating directed k-cliques[9]. Since CPM is an NP-Complete problem, W-CPM is created to lessen the processing time to polynomial time[10]. W-CPM is limited to undirected graphs thus creating a similar algorithm that works for directed graphs will open more possible networks to be classified to their communities.

Chapter 2

Theoretical Framework and Related Literature

2.1 Theoretical Framework

There is no universally accepted definition for a community. A common description for one is that there are more edges inside the community that the rest of the network. A community should have a connectedness inside it such that there is a pair between each vertices and running only inside the cluster. The intra-cluster density, which are the edges within a cluster, should be larger than the average link density, while the inter-cluster density, which are the edges going to the rest of the graph, is much smaller.

The basic composition of a clique is a triangle. The amount of information propagated is dependent on the direction of the edges. There are seven types of triangles. The ones with a 3-cycle provides the most information since it circulates on all three vertices[5].

There are four types of criterion for communities based on social media analysis - complete mutuality, reachability, vertex degree, and comparison of internal and external cohesion. Small subgroups consists mostly of maximal subgraphs or cliques. But in large networks, larger cliques are not frequent and hierarchies

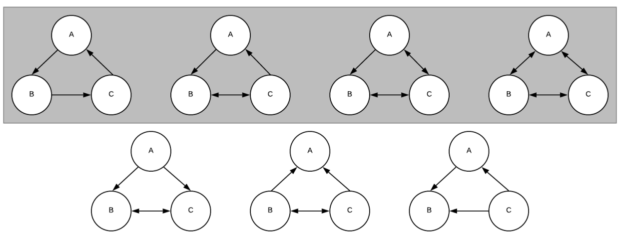

Figure 2.1: The bottom three triangles show little information circulating while the upper four contains 3-cycles

exist. Since a community does not consist of a single large clique, one can define subgroups as clique-like subgraphs. In reachability, one can define a community based on the existence and lengths of paths between vertices. Another criteria is vertex degree, where it roots from the idea that a vertex is adjacent to a minimum number of other vertices in the subgraph. And finally, the comparison of internal and external cohesion, where a strong community is defined based on its intra-cluster and inter-cluster densities such that the former is larger than the latter[3].

CPM is based on complete mutuality criterion of a community in its strictest sense[8]. CPMd redefined a clique such that it takes in consideration the hierarchical structures of networks[9]. W-CPM added a cohesion property by merging triagular cliques by the number of links connected between them and gets the priority of a node based its vertex degree[10].

2.2 Related Work

There are a lot of different algorithms in identifying communities. CPM is rooted to the concept that real-life communities have cliques inside them. To be more precise, CPM defines a k-clique community as a maximal complete subgraph. These cliques are first located, then the communities are identified using a standard component analysis of the clique-clique overlap matrix[8].

Real-time network relationships are not necessarily reciprocal. Thus, G. Palla extended his work on directed networks called CPMd. Here, the k-cliques are replaced by directed k-cliques, defined as complete sub-graphs of size k where it can be any pair of nodes such that there is a directed link pointing from a higher order towards a lower one. For these directed k-cliques, double-links are not allowed, no directed loops and the restricted out degree of each vertex in the clique should differ[9].

Communities do not necessarily form complete cliques. In SCP, a link is created between two nodes to for small k-cliques based on the amount of common neighbors two nodes have and creates an temporary link for these two nodes to create them.[6]. An algorithm proposed by X. Zhang (WCPM) also uses the concept of weak cliques to determine communities, which is based on the smallest cliques formed - triangles.[10]

Computing for the identification of k-cliques and grouping for communities is an NP-complete problem. To solve this, WCPM only gets the weak cliques from two adjacent nodes and these cliques are then merged by priority and similarity to form communities. This solved CPM's problem of getting restricted to k-clique communities while having a computational difficulty of O(dm)[10]. J. Kumpula had a similar algorithm that searches for nodes with at least (k-1)-degrees to get k-cliques and uses merges these cliques in a sequential manner to minimize the time taken in determining the communities.[6]

Chapter 3

Problem Statement

The study aims to create a directed CPM using the concepts provided in W-CPM and have it run in polynomial time. This will aid in classifying large and real-time networks in less processing time compared to CPMd.

This study aims to answer the following questions:

- Is there a way to create a directed CPM using W-CPM concepts?
- Can an algorithm be made such that it is comparable in accuracy and speed on current popular directed network algorithms?

Chapter 4

Objectives of the Study

This study will focus only on finding alternatives on weak cliques and merging mechanisms that will work on directed networks.

The main objectives of this study are as follows:

- Create a directed network method using the main concepts of W-CPM
- Have the algorithm run in polynomial time

Chapter 5

Methodology

5.1 Directed Weak Cliques

Definition 1 (Directed Weak Cliques Determined by an Adjacent Source Node and Drain Node): Given a directed network $G = (V, E)$, where V is the set of nodes and E is a set of links. Let u and v be two adjacent nodes in G with u having a minimum of one out-degree and v having a minimum of one in-degree. A directed weak clique determined by u and v is defined as

$$G_{uv} = (V_{uv}, E_{uv})$$

where $V_{uv} = \{(u, v) \cup (N_u \cap_v)\}$, $E_{uv} = \{(x \rightarrow y) \in E | x, y \in V_{uv}\}$, $N_u = \{x | (x \rightarrow u) \in E\}$, $N_v = \{x | (v \rightarrow x) \in E\}$.

5.2 Node Priority

Definition 2 (Node Strength[2]): Given a weighted network $G = (V, E)$, where V is the set of nodes and E is a set of links. Let u and v be two adjacent nodes in G. The weight of edge e_{uv} is w_{uv} where $w_{uv} = 0$ when nodes u and v are not connected by an edge. The node strength k_u is defined as

$$k_u = \sum_{v \in V} w_{uv}$$

Definition 3 (Common Neighbors[1]):

$$s_{xy} = |N_u \cap N_v|$$

Input: Graph G, Node n Output: number
if G is directed, return number of incoming edges of n if G is weighted, return sum of weights of incoming edges of n

Table 5.1: Pseudocode: Setting the Node Priority

Input: Graph G, Node u, v Output: number
$in_u \leftarrow$ incoming nodes of u $out_v \leftarrow$ outgoing nodes of v return $\lvert in_u \cap out_v \rvert$

Table 5.2: Pseudocode: Getting the Referral Count of a node

Definition 3.5 (Referral Count):

$$s_{xy} = \text{outgoing link nodes} N_u \cup \text{incoming link nodes} N_v$$

5.3 Sequential Algorithm

Definition 4 (SCP Phase II[6]): Given a set of k-cliques, each component (k-1)-cliques are extracted. If the (k-1)-cliques is a component of a different k-clique, then merge. Repeat for the entire set.
Definition 4.5 (Directed Weak Clique Merge): Given a set of directed weak cliques, its directed links are extracted. If the link is a component of a different weak clique, then merge. Repeat for the entire set.

Chapter 6

Testing

6.1 Normalized Mutual Information(NMI)[7]

Chapter 7

Results and Discussion

Chapter 8

Results and Discussion

Chapter 9

Conclusion

Chapter 10

Recommendations

List of References

[1] Bojanowski, M., and Chrol, B. Proximity-based methods for link prediction in graphs with r package linkprediction. https://cran.r-project.org/, 2018. [Online; accessed 27 Feb 2020].

[2] Chen, D., Shang, M., Lv, Z., and Fu, Y. Detecting overlapping communities of weighted networks via a local algorithm. *Physica A: Statistical Mechanics and Its Applications* (2010).

[3] Fortunato, S. Community detection in graphs. *Elsevier Physics Reports* (2009).

[4] Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications* (2018).

[5] Klymko, C., Gleich, D., and Kolda, T. G. Using triangles to improve community detection in directed networks.

[6] Kumpula, J. M., Kivela, M., Kaski, K., and Saramaki, J. Sequential algorithm for fast clique percolation. *Physical Review* (2008).

[7] McDaid, A. F., Greene, D., and Hurley, N. Normalized mutual information to evaluate overlapping community finding algorithms.

[8] Palla, G., Farka, I. J., Derenyi, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature Letters 435* (2005).

[9] Palla, G., Farka, I. J., Pollner, , Derenyi, I., and Vicsek, T. Directed network modules. *New Journal of Physics* (2007).

[10] Zhang, X., Wang, C., Su, Y., Pan, L., and Zhang, H.-F. A fast overlapping community detection algorithm based on weak cliques for large-scale networks. *IEEE Transactions on Computational Social Systems 4* (December 2017).

Chapter 11

Appendix