Discovering Overlapping Community Structure in Social Networks

Z. Bahrami Bidoni
Department of Computer and Information Systems
Clark Atlanta University, Atlanta, GA, USA
Email: z.bahrami@students.cau.edu

Khalil Shujaee
Department of Computer and Information Systems
Clark Atlanta University, Atlanta, GA, USA
Email: kshujaee@cau.edu

Roy George
Department of Computer and Information Systems
Clark Atlanta University, Atlanta, GA, USA
Email: rgeorge@cau.edu

***Abstract*—The massive growth of social networks has created a need for the development of algorithms and systems that can be used for their analysis. Techniques that reveal the structure and the information flow within the network can be used to understand the dynamics of the network and provide new opportunities in promoting virtual communities for a variety of purposes. The basis of this research work is the understanding of a social network community, with special emphasis on communities that overlap. A community is defined as a subgraph with a higher internal density and a lower crossing density with others subgraphs. In this research, we apply a distance based ranking algorithm, the Overlapped Correlation Density based Partitioning (OCDP), to understand communities that overlap. We introduce the OCDP algorithm, and present preliminary results of the technique through its application to a real world data set, the Bottlenose dolphin network. The OCDP is compared with other algorithmic approaches, and in preliminary results show that it has good performance across different evaluation metrics.**

Keywords: Dynamic social network, Organizational structure, Overlapping Community discovery, Correlation Density Rank

I. INTRODUCTION *(HEADING 1)*

Community detection is an significant issue in social network analysis, where the objective is to recognize related sets of members such that intra-community associations are denser than inter-communities associations [2][3][5][6][8]-[11][14][15]. Researchers have presented various methods to extract communities from an SN that paper [17] presented a survey of these studies. Specifically, discovering the organizational structure of communities in an SN has been identified as an interesting but challenging problem [4,13]. Examples of important applications include characterizing potential key candidates for viral marketing or discovering core members of criminal group in monitoring criminal network [13]. Research on finding motivated members in a Social Networks is one component of this researhc, but outcomes have limited power to supply a complete view of the organizational structure.

In the real-world networks, communities are often not disjoint but overlapped to some extent [19]. For example, in social life, a person usually has connections with several social groups such as family, friends, and colleagues; a researcher may collaborate with other researchers in different fields. This can also happen in many other complex networks including biological networks, online social networks, and so on. Indeed, overlap is quite a significant feature in real-world social networks [20]. For this reason, researchers have paid attention to the problem of overlapping community detection, and many techniques have been proposed, such as the the Link method which reinvents communities as groups of links rather than nodes [21], fuzzy c-means clustering [22], and the algorithms utilizing local expansion and optimization including LFM (Local Fitness Maximization) [23], UEOC (the Unfold and Extract Overlapping Communities) [24], DenShrink (Density-based Shrinkage) [25] and the method based on a local definition of community strength [25]. A review of overlapping community detection algorithms is found in [26] along with quality measures and several existing benchmarks. The authors have previously defined the Community Density Rank [18], a measure that is used to evaluate the structure of a community. In this research paper, we extend the CDR algorithm to define the Overlapped Correlation Density based Partitioning (OCDP), to understand communities that overlap, and present initial results from the application of the algorithm to a real world data set, the Bottlenose dolphin network. The OCDP is compared with other algorithmic approaches, and it is shown that it has an equal performance with several published algorithms over a publicly available community data set, the Bottlenose Dolphin Network. It should be noted that this research effort is a work in progress, and though promising the OCDP has to be validated over much larger data sets.

The rest of the paper is organized as follows. Section II introduces the methodology and outlines the OCDP. Section III presents the results of the analysis on a real life data set and Section IV concludes the paper and proposes future work.

II. METHODOLOGY

In the analysis of a network, the first task is to compare nodes. In order to execute this task, the importance of each node within the network has to be understood. The nodes

 ISBN: 978-1-61208-504-3

that link to many other important nodes are themselves important. This process of analysis is similar to PageRank based algorithms [24]. The PageRank algorithm is the best known of these approaches, having been the basis of the original search mechanism for Google. Here the global "importance" ranking for every web page is obtained by analyzing links among web pages. Other algorithms that improve on PageRank such as HITS, OPIC and etc. have been proposed.

The OCDP computation proceeds in two parts- first we compute the Correlation Density Rank (CDR) of each node, and second, we use the CDR to find core nodes and the nodes associated with the cores (the community). The Correlation Density Rank (CDR), is based on finding more frequent and influential Randomized Shortest Paths(RSP)[57] between nodes. In RSP model, the randomness of the walker is constrained by fixing the relative entropy between the distribution over paths according to the reference probabilities and the distribution over paths that the walker actually chooses from. With this constraint, the walker then chooses the path from the probability distribution that minimizes the expected cost. We employ the RSP measurement method in [23] as the distance between nodes, but with one major difference: we consider customized initial cost for edges such that, along with finding shortest path between nodes. The random walker intelligently selects the most important neighbor resulting in lower cost and smaller distance. The CDR considers the distance between nodes as punishment and computes the density ranks of nodes. Hence, there will be a larger traffic amongst shortest path of nodes, if the distance becomes smaller. If the distance between nodes, *i* and *j* was less than the distance between *i* and *k*, then, *i*'s rank effect on *j* is more than on *k*, and the probability that a random surfer reaches *j* from *i* is more than the probability to reach *k*. Therefore, the objective is to minimize punishment so that a node with less distance entropy to have a higher rank. The CDR scores of a node are compared with the nodes in its vertex border to determine the "core" of the community. Communities are then constructed around the cores iteratively, using a membership formulation, where each node can participate with communities formed by multiple cores.

Definition 1 (Cardinality of a community). The cardinality of a community C is the number of its vertices. It is denoted by $|C|$.

Definition 2 (Direct neighbor). In the graph *G = (V, E)*, the vertex v is a direct neighbor of the node u if v and u are connected by an edge. This relationship is represented by the edge $(v,u) \in E$.

Definition 3 (Vertex border). It is all the direct neighbors of node v in the graph. This set is noted by B(v). More formally this quantity is noted as follows:

$$B(v) = \{u \in V; \ \{u,v\} \in E\}$$

Definition 4 (Internal Degree of a vertex to a community). We call internal degree of a vertex v to a community C as the number of edges that point towards members of C**.**

$$d_{in}(v,C) = \left|\{(v,v') \in E, \ \ v' \in C\}\right|$$

Definition 5 (External Degree of a vertex to a community). We call external degree of a node v to a community C as the number of its direct neighbors who are not in C**.**

$$d_{ext}(v,C) = \left|\{(v,v') \in E, \ \ v' \notin C\}\right|$$

Definition 6 (Average distance between a node and a community). It is the sum of distances of node u to different nodes $v \in C$, divided by the cardinality of C**.**

$$dist_{avrage}(u,C) = \begin{cases} \dfrac{\sum_{v \in C} RSP(u,v)}{|C-1|} & if \ \ u \in C \\ \dfrac{\sum_{v \in C} RSP(u,v)}{|C|} & otherwise \end{cases}$$

Definition 7 (Weighting coefficient). It is the degree of compactness of one node u to a community C**.**

$$\rho(u,C) = \frac{|B(u)|}{d_{in}(u,C)}$$

Definition 8 (Membership degree). The membership degree of node v to community C is given by:

$$Md(u,C) = \frac{1}{dist_{avrage}(u,C) * \rho(u,C)}$$

Definition 9 (Influence Cofficient degree) where λ is the parameter of control overlapping extent of communities.

$$F_C^u = 2\frac{\lambda * dist_{in}^u - (1-\lambda) * dist_{ext}^u}{dist_{in}^u + dist_{ext}^u}$$

Algorithm 1. Calculating m-Score for members: Correlation Density Rank (CDR)

Input: social network G

Out: vector of m-Score for all members R

1. Initialize cost distance matrix C

$$C[i, j] = \log_{\left(1-w_{ij}^{in} w_{ij}^{out}\right)}^{(1-\exp(-\gamma f_{ij}))}$$

2. Finding the matrix of RSP dissimilarities [43]: {

a. W ← $P^{ref} \circ \exp(-\beta C)$

b. $Z \leftarrow (I - W)^{-1}$

(Note that $(I - W)^{-1} \approx I + W + W^2 + W^3 + \ldots$)

c. $S \leftarrow (Z(C \circ W)Z) \div (Z + \varepsilon)$

d. $\tilde{C} \leftarrow S - ed_s^T$

e. $\Delta^{RSP} \leftarrow \lambda\tilde{C} + (1-\lambda)\tilde{C}^T \qquad 0 \le \lambda \le 1$

}

3. M ← Normalize matrix Δ^{RSP} on rows
4. For each node n_i ($1 \le i \le k$) compute the entropy of related row from matrix M:

a. $E_i \leftarrow -\frac{1}{Lnk}\sum_{j=1}^{k} M_{ij} Ln(M_{ij})$

b. $d_i \leftarrow 1 - E_i$

c. $R_i \leftarrow d_i \Big/ \sum_{i=1}^{k} d_i$

5. Return R

Algorithm 2: Overlapped Correlation Density based Partitioning (OCDP)

Data: A graph $G = (V, E)$

Begin

1: Calculate Correlation Density Rank of all nodes (see Algorithm 1)

2: *u*, if *CDR(u)* > *CDR(B(u))* → *u* is core of the Community

3: For all cores do extend algorithm {

Build border of C: $edg(C) = \{v_i | v_i \in B(C)\}$.

While $(edg(C) \ne \varnothing)$ do

Choose the candidate node v_i of $edg(C)$ which has the highest membership degree to C.

If $F_C^{v_i} > 0$ *then*

$C \leftarrow \{C\} \cup \{v_i\}$

Update of $edg(C)$

else

$edg(C) \leftarrow \varnothing$

end

End

Return C

End.

III. Results

An experimental analysis of OCDP using a publicly available data set is described. We compared OCDP with five well-known algorithms: (1) CFinder (CPM) which implements the clique percolation (2011); (2) COPRA which is based on label propagation (2010); (3) GCE greedy approach (2013); and (4) EAGLE modularity-based approach (Eagle Community Detection Algorithm, 2012). (5) DOCNet (2014). Bottlenose dolphin network is the real and well-known Dolphins social network which describes the associations between 62 dolphins living in Doubtful Sound, New Zealand (Figure 1). The relationship between dolphins represent the statistically significant frequent association between them. This network is interesting because, during the course of the study, the dolphin group split into three smaller subgroups following the departure of key members of the population. In four commonly used measures in the overlapping community structure research, the modularity, Q_{ov} ; the M rank; number of detected overlapping nod es O_n^d and detected memberships O_m^d,) the OCDP had similar or better results (Table 1). The measure evaluations are as follows (indicates better performance): Q_{ov}, O_n^d, O_m^d: higher, M: lower. While the results of the OCDP in comparison to other published techniques looks promising, it should be noted that this is a research effort in progress.

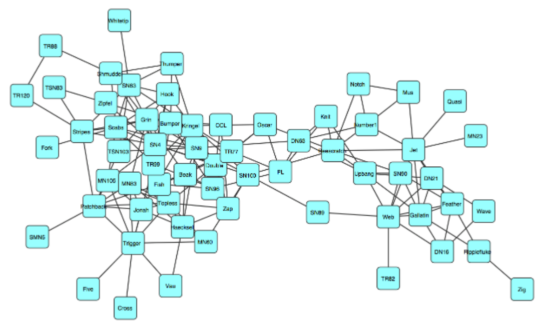

Figure 1. Bottlenose dolphin network.

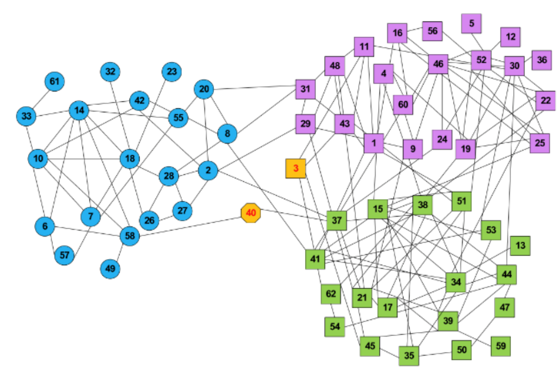

Figure 2. Detected overlapped Communities in Dolphin Network

TABLE I. QUALITY MEASURE COMPARISON

	COPRA (2010)	CPM (2011)	EAGLE (2012)	GCE (2013)	DOC-NET (2014)	OCDP (2015)
Q_{ov}	0.32	0.29	0.32	0.33	0.41	0.47
M	3.00	4.00	4.00	4.00	3.00	3.00
O_m^d	2.00	2.00	2.00	2.00	2.00	2.00
O_n^d	1.75	2.00	1.50	2.00	1.66	2.00

IV. CONCLUSION

Social networks have become an ubiquitous feature of a highly connected global network of users. Analysis of these networks is difficult due to the massive scale of the network and the complexity of the connectivity. Of special interest is the structure and the information flow within the network. Knowledge of these may be leveraged to provide a basis for virtual communities that interact to achieve common goals in a number of domains. In this research, we developed an algorithm the Overlapped Correlation Density based Partitioning (OCDP), that attempts to understand the structure of communities that share members. We present preliminary results of the OCDP technique through its application to a real world data set, the Bottlenose dolphin network. The Dolphin network while interesting is somewhat limited in the number of participants and their interactions. Currently popular social networks involve hundreds of millions of participants, with billions of interactions and the scale up of this technique needs to be investigated.

ACKNOWLEDGMENT

This research is funded in part by the Department of Energy under Contract Number DE-NA 0002686. Any opinions, findings, conclusions or recommendations expressed here are those of the author(s) and do not necessarily reflect the views of the sponsor.

REFERENCES

[1] Kathleen M. Carley, Jana Diesner, Jeffrey Reminga, Maksim Tsvetovat, “Toward an interoperable dynamic network analysis toolkit, Decision Support Systems,” 43 (2007) 1324–1347. 1

[2] C. Chekuri, A. Goldberg, D. Karger, M. Levin, C. Stein, “Experimental study of minimum cut algorithms,” The Proceedings of the 8th SAIM Symposium on Discreet Algorithm, 1997, pp. 324–333. 2

[3] C. Ding, X. He, H. Zha, M. Gu, H. Simon, “A min–max cut algorithm for graph partitioning and data clustering,” The Proceedings of the 2001 IEEE International Conference on Data Mining, 2001, pp. 107–114. 3

[4] Amit Goyal, Francesco Bonchi, Laks V.S. Lakshmanan, “Discovering leaders from community actions,” The Proceedings of 17th ACM conference on Information and knowledge management, 2008, pp. 499–508. 4

[5] L. Hagen, A.B. Kahng, “New spectral methods for ratio cut partitioning and clustering,” IEEE Transactions on Computed Aided Design 11 (9) (1992) 1074–1085. 6

[6] Bo Long, Xiaoyun Wu, Zhongfei (Mark) Zhang, “Community learning by graph approximation,” The proceedings of 7th IEEE International Conference on Data Mining, 2007, pp. 232–241. 7

[7] Hao Ma, Haixuan Yang, Michael R. Lyu, Irwin King, “Mining social networks using heat diffusion processes for marketing candidates selection,” The proceedings of 17th ACM conference on Information and knowledge management, 2008, pp. 233–242.8

[8] M.E.J. Newman, “Fast algorithm for detecting community structure in networks,” Physical Review E 69 (2004) 066133.9

[9] M.E.J. Newman, M. Girvan, “Finding and evaluating community structure in networks,” Physical Review E 69 (2) (2004) 1–15.10

[10] J. Shi, J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.11

[11] Andrew Y. Wu, et al., “Mining scale-free networks using geodesic clustering,” The Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 719–724. 12

[12] Jennifer J. Xu, Hsinchun Chen, “Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks,” Decision Support Systems 38 (2004) 473–487.13

[13] J. Jennifer Xu, Hsinchun Chen, “CrimeNet explorer: a framework for criminal network knowledge discovery,” ACM Transactions on Information Systems 23 (2) (2005) 201–226.14

[14] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, Thomas A.J. Schweiger, “SCAN: a structural clustering algorithm for networks,” The Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 824–833.15

[15] Ding Zhou, Eren Manavoglu1, Jia Li, C. Lee Giles, Zha Hongyuan, “Probabilistic models for discovering e-communities,” The Proceedings of the 15th international conference on World Wide Web, 2006, pp. 173–182.16

[16] Kivimäki, Ilkka, Masashi Shimbo, and Marco Saerens. "Developments in the theory of randomized shortest paths with a comparison of graph node distances." Physica A: Statistical Mechanics and its Applications 393 (2014): 600-616.17

[17] Malliaros, F. D., & Vazirgiannis, M. (2013). “Clustering and community detection in directed networks: A survey.” Physics Reports, 533(4), 95-142.18

[18] Z. Bahrami Bidoni, R.George, “Discovering Community Structure in Dynamic Social Networks using the Correlation Density Rank,” in SocialCom - Stanford, CA, USA. The Sixth ASE International Conference on Social Computing, 2014 19

[19] G. Palla, I. Derényi, I. Farkas, T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature 435 (2005) 814–818.20

[20] S. Kelley, M.K. Goldberg, K. Mertsalov, M. Magdon-Ismail, W. Wallace, “Overlapping communities in social networks,” Int. J. Social Comput. Cyber. Phys. Syst. 1 (2) (2011) 135–159.21

[21] S. Zhang, R.S. Wang, X.S. Zhang, “Identification of overlapping community structure in complex networks using fuzzy c-means clustering,” Physica A 374 (2007) 483–490.22

[22] I. Psorakis, S. Roberts, M. Ebden, B. Sheldon, “Overlapping community detection using Bayesian non-negative matrix factorization,” Phys. Rev. E 83 (2011) 066114.23

[23] A. Lancichinetti, S. Fortunato, J. Kertész, “Detecting the overlapping and hierarchical community structure in complex networks,” New J. Phys. 11 (2009) 033015.24

[24] D. Jin, B. Yang, C. Baquero, D. Liu, D. He, J. Liu, “A markov random walk under constraint for discovering overlapping communities in complex networks,” J. Stat. Mech: Theory. E 2011.05 (2011) P05031.25

[25] J.B. Huang, H.L. Sun, J.W. Han, B.Q. Feng, “Density-based shrinkage for revealing hierarchical and overlapping community structure in networks,” Physica A 390 (2011) 2160–2171.26

[26] J. Xie, S. Kelley, B.K. Szymanski, “Overlapping community detection in networks: the state of the art and comparative study,” ACM. Comput. Surv. (2013) Article No. 43.27

 ISBN: 978-1-61208-504-3