An Empirical Study of Community Overlap: Ground-truth, Algorithmic Solutions, and Implications

Joyce Jiyoung Whang
Sungkyunkwan University
Suwon, Korea
jjwhang@skku.edu

ABSTRACT

In real-world social networks, communities tend to be overlapped with each other because a vertex can belong to multiple communities. To identify these overlapping communities, a number of overlapping community detection methods have been proposed over the recent years. However, there have been very few studies on the characteristics and the implications of the community overlap. In this paper, we investigate the properties of the nodes and the edges placed within the overlapped regions between the communities using the ground-truth communities as well as algorithmic communities derived from the state-of-the-art overlapping community detection methods. We find that the overlapped nodes and the overlapped edges play different roles from the ones that are not in the overlapped regions. Using real-world data, we empirically show that the highly overlapped nodes are involved in structure holes of a network. Also, we show that the overlapped nodes and edges play an important role in forming new links in evolving networks and diffusing information through a network.

KEYWORDS

community detection; overlap; social network analysis

1 INTRODUCTION

A social network can be represented as a graph where individuals are denoted by a set of vertices and the social relationships between the individuals are denoted by a set of edges of the graph. Community detection is one of the most important and fundamental tasks in social network analysis where the goal is to identify a set of cohesive nodes that are densely connected with each other. Since an individual usually participates in more than one social circle, the communities naturally overlap with each other.

Unlike the traditional graph clustering problem where a graph is partitioned into disjoint clusters, there exist overlapped regions between communities in the overlapping community detection problem. Intuitively, we can expect that the nodes and the edges placed within the overlapped regions may play different roles from the ones that are not in the overlapped regions.

CIKM'17, November 6–10, 2017, Singapore.

ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3133133

Table 1: Summary of real-world networks.

Graph	No. of vertices	No. of edges	Ground-truth
DBLP	317,080	1,049,866	✓
LiveJournal	1,143,395	16,880,773	✓
Flickr-a	1,994,422	21,445,057	N/A
Myspace-a	2,086,141	45,459,079	N/A
LiveJournal-a	1,757,326	42,183,338	N/A

Table 2: Ground-truth Communities.

	DBLP	LiveJournal
No. of communities	13,477	662,859
No. of overlapped nodes (%)	110,806 (35%)	752,537 (65%)
No. of overlapped edges (%)	356,801 (34%)	4,724,058 (28%)

In this paper, we investigate the characteristics and the implications of the overlapped nodes and the overlapped edges based on the ground-truth overlapping communities as well as algorithmic overlapping communities. In particular, we find that the highly overlapped nodes bridge different communities and might comprise a part of structural holes in a network. Also, we study the properties of a set of newly formed edges in evolving networks finding that the new links tend to be formed in the overlapped regions. Finally, we implement a simple information diffusion model based a networked coordination game, and show that the overlapped nodes and the overlapped edges are crucial in information spreading.

2 DEFINITIONS

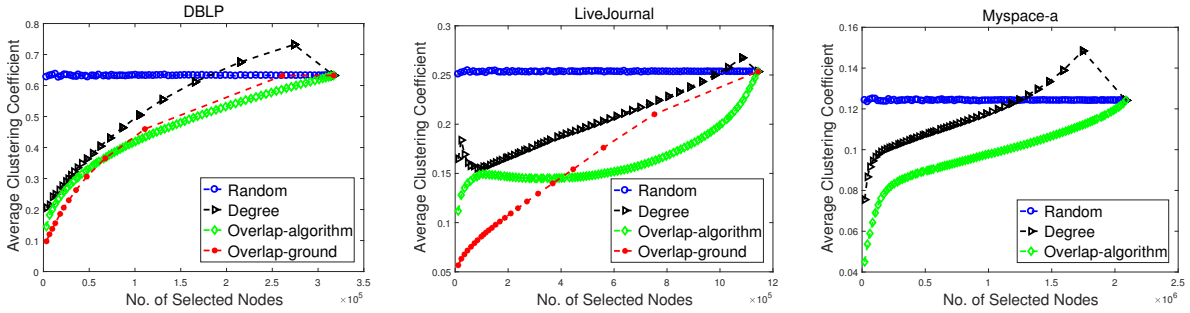
Given a graph $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ indicates a set of vertices and $\mathcal{E}$ indicates a set of edges, an overlapping community detection algorithm finds communities that are not necessarily pairwise disjoint. That is, a vertex is allowed to belong to multiple communities.

For each vertex $v_i \in \mathcal{V}$ ($i = 1, \cdots, n$ where $n = |\mathcal{V}|$), let $\mathcal{S}_i$ denote a set of communities the vertex v_i belongs to. We assume that graphs are undirected. We define the *overlapped nodes* and the *overlapped edges* as follows.

Definition 1 (Overlapped nodes). *We say that "a vertex v_i is placed in an overlapped region" or "a vertex v_i is an overlapped node" if the vertex belongs to more than one community, i.e., $|\mathcal{S}_i| \geq 2$.*

Definition 2 (Overlapped edges). *We say that "an edge $e = \{v_i, v_j\}$ is placed in an overlapped region" or "an edge e is an overlapped edge" if $|\mathcal{S}_i \cap \mathcal{S}_j| \geq 2$.*

Let $\mathcal{V}_\tau$ denote the set of overlapped nodes of a graph $G = (\mathcal{V}, \mathcal{E})$, and $\mathcal{E}_\tau$ denote the set of overlapped edges. Then, we define the set of *non-overlapped nodes* as $\mathcal{V} \setminus \mathcal{V}_\tau$. Similarly, we define the set of *non-overlapped edges* as $\mathcal{E} \setminus \mathcal{E}_\tau$.



(a) DBLP (b) LiveJournal (c) Myspace-a

Figure 1: The average clustering coefficients. Highly-overlapped nodes tend to have low clustering coefficients.

3 EXPERIMENTAL SETUP

We use five real-world networks from [5], [8], [10]. Table 1 shows a summary of the graphs. From [5], we get the ground-truth communities for the DBLP and LiveJournal datasets which are well-known benchmarks for overlapping community detection (see Table 2 for details). In particular, these datasets have been studied in the context of the community-affiliation graph model [13] which shows that the community overlap consists of densely connected nodes, which might be also related to our investigation in Section 5.

Among a number of overlapping community detection algorithms [12], a local expansion approach has been known to be one of the most successful and scalable methods. For example, it has been shown that the personalized PageRank-based local expansion methods [1] can produce communities that are similar to the ground-truth communities [4]. In particular, a recently proposed method called NISE [11] can efficiently identify high-quality overlapping communities using a PageRank-based local expansion from a set of good seed nodes in a graph. Since we notice that the NISE method achieves the best accuracy among the state-of-the-art overlapping community detection methods (see [11] for details), we focus on the algorithmic solutions produced by the NISE method in this paper. Within the NISE method, we try the different seeding and expansion methods, and select the one that shows the most similar patterns to the ground-truth communities.

4 OVERLAPPED NODES AND STRUCTURAL HOLES IN A NETWORK

Each node in a graph might have different characteristics based on its position in the graph. For example, some nodes might be placed in a center of a tightly-knit community whereas some nodes might be placed in the boundaries of a community. A subset of the boundary nodes might comprise a *structural hole* in a network [2]. The structural hole is defined to be an empty space of a network between two sets of nodes that do not closely interact with each other. This space is composed of a set of nodes that have multiple *local bridges* (see [3] for more details). Note that if a node is adjacent to many local bridges, then the node has a low clustering coefficient[1].

In Figure 1, we show an important characteristic of highly overlapped nodes in terms of their clustering coefficients. Let us define the degree of overlap (or the *overlap degree*) of a node as the number of communities the node belongs to (i.e., $|S_i|$). Then, we sort the nodes according to their overlap degrees in descending order. Let t_p denote the overlap degree of the $\lceil pn \rceil$-th node ($0 \leq p \leq 1$) where n is the total number of nodes. We select the nodes whose overlap degrees are greater than or equal to t_p, and compute their average clustering coefficient. The x-axis of the plots indicates $\lceil pn \rceil$ as we increase p and the y-axis represents the average clustering coefficient. 'Overlap-ground' indicates that we compute the overlap degree based on the ground-truth communities whereas 'Overlap-algorithm' is based on the algorithmic communities. For comparison, we also select $\lceil pn \rceil$ nodes by selecting top $\lceil pn \rceil$ nodes according to the degree centrality (i.e., the number of neighbors). This is denoted by 'Degree' in the plots. The 'Random' line indicates the case where we randomly select $\lceil pn \rceil$ nodes, and thus, the line corresponds to the average clustering coefficient of the entire nodes. In Figure 1, we see that high-overlap nodes tend to have low clustering coefficients. As the overlap degree increases, the average clustering coefficient decreases (note that we interpret the plots from right to left; the threshold of the overlap degree decreases from left to right of the x-axis). It is interesting to see that high-overlap nodes have even lower clustering coefficients than high-degree nodes. Since the denominator of the clustering coefficient of a vertex v_i is defined to be $d_i(d_i-1)/2$ where d_i is the degree of v_i, it is likely that high-degree nodes have low clustering coefficients. Nonetheless, the 'Overlap-*' line is even lower than the 'Degree' line in Figure 1. Nodes with low clustering coefficients indicate that those nodes have diverse neighbors who are not directly connected with each other (i.e., they might be adjacent to multiple local bridges). Thus, we can infer that the highly-overlapped nodes might *bridge* different communities, and play as structural holes in a network. This observation is also consistent with [7] even though our empirical analysis provides different viewpoints from [7].

Table 3: Networks with Timestamps.

Graph	# of vertices	# of *new* edges
Flickr-a ⟶ Flickr-b	1,994,422	395,880
Myspace-a ⟶ Myspace-b	2,086,141	334,679
LiveJournal-a ⟶ LiveJournal-b	1,757,326	649,909

5 NEW LINKS IN COMMUNITY OVERLAP

Social networks keep changing over time, e.g., new links are formed over time. Link prediction is an important task in social network analysis where the goal is to predict a set of new edges that are likely to be formed in the near future. We investigate the patterns

[1]The clustering coefficient of a vertex v_i is defined to be the probability that two randomly selected neighbors of v_i are directly connected with each other.

Table 4: Classification of the edges according to the number of common communities of the endpoints of the edges.

	Flickr-b		LiveJournal-b		Myspace-b	
	Ground ($\mathcal{Q}$)	Random ($\mathcal{R}$)	Ground ($\mathcal{Q}$)	Random ($\mathcal{R}$)	Ground ($\mathcal{Q}$)	Random ($\mathcal{R}$)
$\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert = 0$	73,858 (18.66%)	223,995 (56.58%)	8,940 (1.38%)	402,832 (61.98%)	1,159 (0.35%)	62,047 (18.54%)
$\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert = 1$	64,112 (16.19%)	103,164 (26.06%)	6,290 (0.97%)	99,433 (15.30%)	2,219 (0.66%)	48,372 (14.45%)
$\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert \geq 2$	257,910 (65.15%)	68,721 (17.36%)	634,679 (97.66%)	147,644 (22.72%)	331,301 (98.99%)	224,260 (67.01%)
mean($\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert$)	4.77	0.68	20.00	1.23	26.15	5.03
median($\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert$)	3	0	15	0	20	3

Table 5: The overlap degrees of the endpoints of the links.

	Flickr-b		LiveJournal-b		Myspace-b	
	$\mathcal{Q}$	$\mathcal{R}$	$\mathcal{Q}$	$\mathcal{R}$	$\mathcal{Q}$	$\mathcal{R}$
mean	20.6	12.1	68.8	45.8	146.0	87.2
median	16	10	58	41	103	80

of the link formations in the overlapped regions of a network using three real-world datasets summarized in Table 3. In these datasets, we have the information about a set of new links formed during a month for a fixed set of vertices. Let $G_a = (\mathcal{V}_a, \mathcal{E}_a)$ denote a graph at time t and $G_b = (\mathcal{V}_b, \mathcal{E}_b)$ denote a graph at time $t+1$ where $\mathcal{V}_a \equiv \mathcal{V}_b$ and $\mathcal{E}_a \subset \mathcal{E}_b$. For example, Flickr-a corresponds to G_a and Flickr-b corresponds to G_b for our datasets. Then, the set of new links denoted by $\mathcal{Q}$ can be represented as $\mathcal{Q} = \mathcal{E}_b \setminus \mathcal{E}_a$. Let m denote the number of new links, i.e., $|\mathcal{Q}| = m$.

It has been known that there exists an underlying mechanism that drives to forming a new link between nodes [3]. To examine the characteristics of the new links, we construct a baseline edge set $\mathcal{R}$ by randomly selecting m edges such that $\mathcal{R} \cap \mathcal{E}_a = \emptyset$, i.e., $\mathcal{R}$ does not contain an existing edge at time t. Table 5 shows the mean and the median value of the overlap degrees of the endpoints of the edges in $\mathcal{Q}$ and $\mathcal{R}$. We see that the overlap degrees are greater in $\mathcal{Q}$ than $\mathcal{R}$. This indicates that the new links are likely to be formed around high-overlap nodes. Now, we classify the m edges in $\mathcal{Q}$ according to the community information of the endpoints of the edges. Given an edge $\{v_i, v_j\}$, let us define $\mathcal{X} := \{\{v_i, v_j\} \in \mathcal{Q} : \mathcal{S}_i \cap \mathcal{S}_j = \emptyset\}$. Also, we define $\mathcal{Y} := \{\{v_i, v_j\} \in \mathcal{Q} : |\mathcal{S}_i \cap \mathcal{S}_j| = 1\}$ and $\mathcal{Z} := \{\{v_i, v_j\} \in \mathcal{Q} : |\mathcal{S}_i \cap \mathcal{S}_j| \geq 2\}$. Then, $\mathcal{Q} = \mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}$. Note that $\mathcal{X}$ indicates a set of between-community edges, $\mathcal{Y}$ indicates a set of non-overlapped within-community edges, and $\mathcal{Z}$ indicates the overlapped edges. We similarly decompose $\mathcal{R}$ into the three sets, and Table 4 shows the results. We see that most of the new links are formed within communities by observing that $|\mathcal{X}| \ll |\mathcal{Y} \cup \mathcal{Z}|$ for $\mathcal{Q}$. That is, a new link is likely to be formed between two nodes that belong to the same community. More importantly, we notice that $\mathcal{Q}$ contains a significant number of overlapped edges compared to $\mathcal{R}$. This indicates that the new links are formed in the overlapped regions, i.e., the new links tend to be overlapped edges. Also, when we compare the number of common communities of the endpoints of the links in $\mathcal{Q}$ and $\mathcal{R}$ (the last two rows of Table 4), we see that the new edges include highly overlapped edges.

6 INFORMATION DIFFUSION THROUGH OVERLAPPED NODES AND EDGES

Information diffusion is another important task in social network analysis where the goal is to model the way how information is propagated throughout the network. It has been recognized that a community structure affects the patterns of information spreading [6], [9]. However, most of the information propagation models assume disjoint communities rather than overlapping communities.

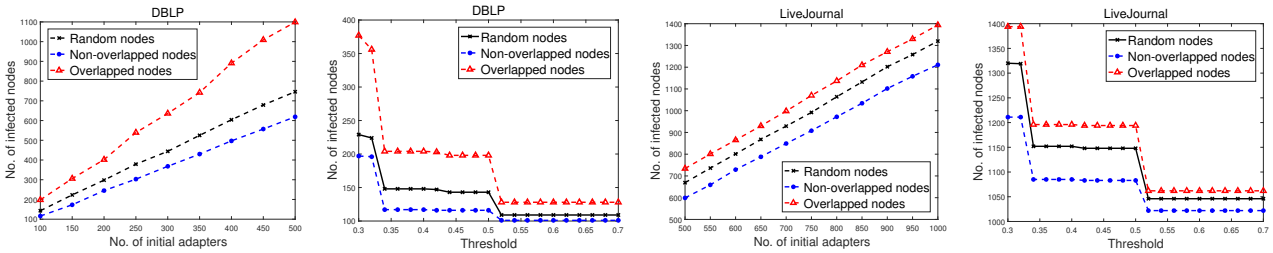
Algorithm 1 Information diffusion based on a coordination game

Input: graph $G = (\mathcal{V}, \mathcal{E})$, a set of initial nodes $\mathcal{V}_0$, a threshold q
Output: a set of infected nodes $\mathcal{V}_I$

```
V_I = V_0, V_T = V_0.
while V_I ≠ V and V_T ≠ ∅ do
    for each v_i ∈ V_T do
        V_T = V_T \ {v_i}.
        if at least q fraction of v_i's neighbors are in V_I then
            V_I = V_I ∪ {v_i}.
            for each v_j such that {v_i, v_j} ∈ E do
                if v_j ∉ V_I then
                    V_T = V_T ∪ {v_j}.
                end if
            end for
        end if
    end for
end while
```

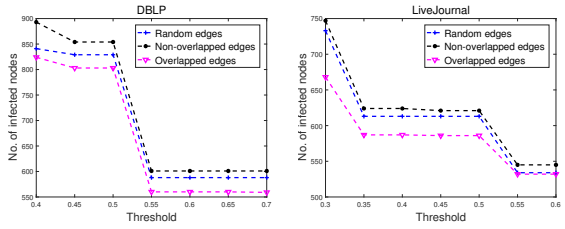
We explore the importance of overlapped nodes and overlapped edges by considering a simple information diffusion model based on a networked coordination game [3]. In this model, it is assumed that each node has a choice between two possible behaviors A and B, and decides to adopt one of the behaviors based on the choices of its neighbors. If there exists an edge between v_i and v_j and the nodes decide to choose the same behavior, there is an incentive for them. This can be represented as a coordination game as follows. Let $a > 0$ denote the payoff if v_i and v_j both decide to adopt the behavior A. Similarly, let $b > 0$ denote the payoff if they adopt the behavior B. There is no payoff if v_i and v_j decide to adopt different behaviors. In this setting, each node v_i in the network chooses A or B so that the node can maximize its payoff. Suppose that v_i has d_i neighbors and p fraction of its neighbors adopt A. Then, v_i should choose to adopt A if $apd_i \geq b(1-p)d_i$. Thus, we get $p \geq b/(a+b)$. Let $q = b/(a+b)$ denote the threshold. We see that v_i chooses A if q fraction of its neighbors choose A. Note that if the payoff a is significantly larger than b, then a node chooses to adopt A even though only a small fraction of its neighbors chooses A. See [3] for more details. We apply this simple information diffusion model to our networks. We assume that there is a set of initial nodes $\mathcal{V}_0$ that adopt A while the rest of the nodes in the network adopt B. Then, each node plays the coordination game, and we see which nodes end up deciding to adopt A. Algorithm 1 describes the procedure.

To investigate the roles of the overlapped nodes in this information diffusion process, we choose the initial node set $\mathcal{V}_0$ in three



(a) Varying $|\mathcal{V}_0|$ on DBLP (b) Varying q on DBLP (c) Varying $|\mathcal{V}_0|$ on LiveJournal (d) Varying q on LiveJournal

Figure 2: Information diffusion with different initial nodes. Overlapped nodes play a crucial role in information spreading.



(a) Varying q on DBLP (b) Varying q on LiveJournal

Figure 3: Information diffusion with differently removed edges. Overlapped edges are important in diffusion.

different ways: (i) random nodes, (ii) non-overlapped nodes, and (iii) overlapped nodes. Then, we count the number of infected nodes in the network, i.e., the number of nodes that decide to adopt A. This corresponds to the size of the output $\mathcal{V}_I$ of Algorithm 1. In Figure 2, 'Random nodes' indicates the case where we randomly choose k nodes where k is the number of initial adopters (i.e., $|\mathcal{V}_0| = k$). 'Non-overlapped nodes' indicates that we randomly select k initial adopters among the non-overlapped nodes in the network whereas 'Overlapped nodes' indicates we randomly select k nodes among the overlapped nodes (we use the ground-truth communities). In Figure 2, we analyze the number of infected nodes for each of the three different initial node sets with different numbers of initial adopters k and different threshold values q. We repeat the experiments ten times and show representative plots in Figure 2. We see that the number of infected nodes is maximized when we select the initial nodes among the overlapped nodes. Also, if we construct $\mathcal{V}_0$ using the non-overlapped nodes, the number of infected nodes is even less than the random selection. This indicates that whether a node is an overlapped node or not is an important factor to determine the success of information spreading, and the overlapped nodes tend to effectively spread the information through the network.

To investigate the roles of overlapped edges in information diffusion, we now remove m_r edges in the network in three ways: (i) random edges, (ii) non-overlapped edges, and (iii) overlapped edges. That is, we remove m_r edges randomly, or remove m_r edges among non-overlapped edges, or remove m_r edges among overlapped edges. We use the ground-truth communities, and set $m_r = 0.25m$ where m is the total number of edges in the network (i.e., we remove a quarter of the edges). Given a fixed set of randomly chosen initial adopters, we count the number of infected nodes in Figure 3 by varying the threshold value q. We see that the information is not spread well when the overlapped edges are removed. Thus, we can infer that the overlapped edges have much contributions to information diffusion than the randomly selected edges and the non-overlapped edges. All these results imply that the overlapped nodes and the overlapped edges are crucial in information propagation through the network.

7 CONCLUSIONS

We analyze various characteristics of the overlapped nodes and the overlapped edges by conducting empirical studies on the ground-truth and algorithmic overlapping communities. We show that high-overlap nodes have low clustering coefficients–they *bridge* different communities, which indicates that they might play as structural holes in a network. Also, we find that when networks evolve over time, the new links tend to be formed within overlapped regions of the graph. Finally, we observe that the overlapped nodes and the overlapped edges play a critical role in spreading information throughout the network. We expect that our investigations can provide useful intuition and insight for many practical applications including link prediction and information propagation models.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the NRF of Korea funded by MOE(2016R1D1A1B03934766) and by the National Program for Excellence in SW supervised by the IITP(2015-0-00914), Korea.

REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. 2006. Local Graph Partitioning using PageRank Vectors. In *FOCS*. 475–486.
[2] R. S. Burt. 2004. Structural Holes and Good Ideas. In *AJS*, Vol. 110.
[3] D. Easley and J. Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press.
[4] I. M. Kloumann and J. M. Kleinberg. 2014. Community Membership Identification from Small Seed Sets. In *KDD*. 1366–1375.
[5] J. Leskovec and A. Krevl. 2014. Stanford Network Analysis Project. http://snap.stanford.edu/data. (2014).
[6] S. Lin, Q. Hu, G. Wang, and P. S. Yu. 2015. Understanding Community Effects on Information Diffusion. In *PAKDD*. 82–95.
[7] T. Lou and J. Tang. 2013. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In *WWW*. 825–836.
[8] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. 2008. Growth of the Flickr Social Network. In *WOSN*. 25–30.
[9] Stephen Morris. 2000. Contagion. In *The Review of Economic Studies*, Vol. 67.
[10] H. H. Song, B. Savas, T. W. Cho, V. Dave, Z. Lu, I. S. Dhillon, Y. Zhang, and L. Qiu. 2012. Clustered Embedding of Massive Social Networks. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40. 331–342.
[11] J. J. Whang, D. F. Gleich, and I. S. Dhillon. 2016. Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion. In *TKDE*, Vol. 28.
[12] J. Xie, S. Kelley, and B. K. Szymanski. 2013. Overlapping Community Detection in Networks: the State of the Art and Comparative Study. In *CSUR*.
[13] J. Yang and J. Leskovec. 2014. Structure and Overlaps of Ground-Truth Communities in Networks. In *TIST*, Vol. 5. 26:1–26:35.