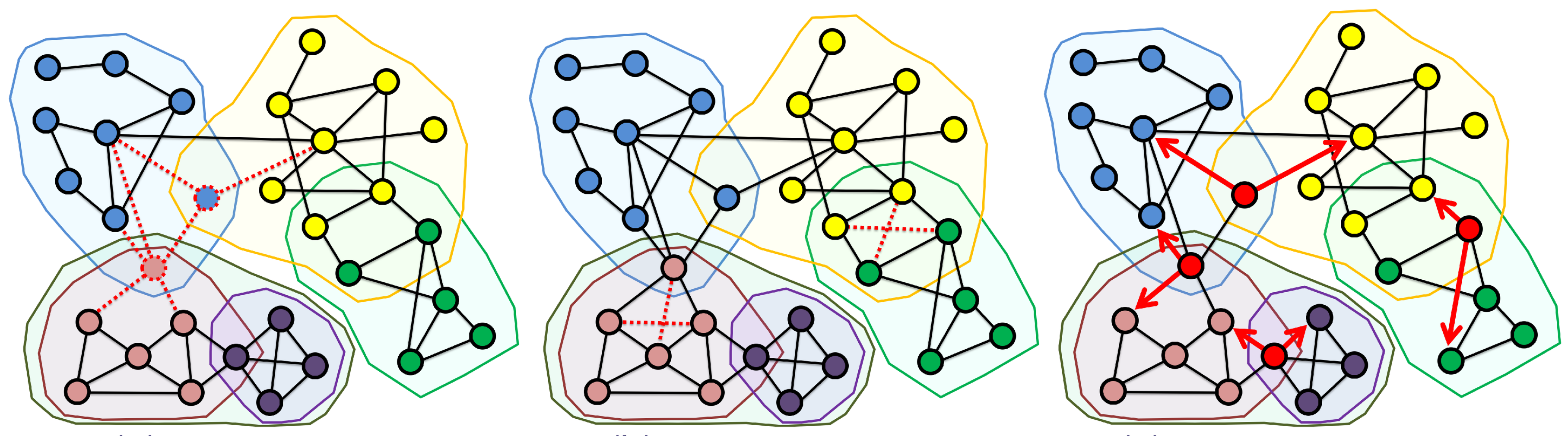
An Empirical Study of Community Overlap: Ground-truth, Algorithmic Solutions, and Implications

Joyce Jiyoung Whang
Sungkyunkwan University (SKKU)
ACM International Conference on Information and Knowledge Management, 2017

Main Contributions

- We investigate the properties of the nodes and the edges placed within the overlapped regions between different communities.
- Overlapped nodes and overlapped edges play different roles from the ones that are not in the overlapped regions.
- Highly overlapped nodes are involved in structure holes of a network.
- Overlapped nodes and edges play an important role in forming new links and diffusing information through a network.



(a) Structural Holes (b) Forming New Links (c) Information Diffusion

Definitions & Experimental Setup

- Let $\mathcal{S}_i$ denote a set of communities a vertex v_i belongs to.
  - A vertex v_i is an overlapped node if $|\mathcal{S}_i| \geq 2$.
  - An edge $e = \{v_i, v_j\}$ is an overlapped edge if $|\mathcal{S}_i \cap \mathcal{S}_j| \geq 2$.
- We have the ground-truth communities for DBLB and LiveJournal.
- We use the NISE method* to produce algorithmic communities.

(* J. Whang et al., "Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion", *TKDE*, 2016.)

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.

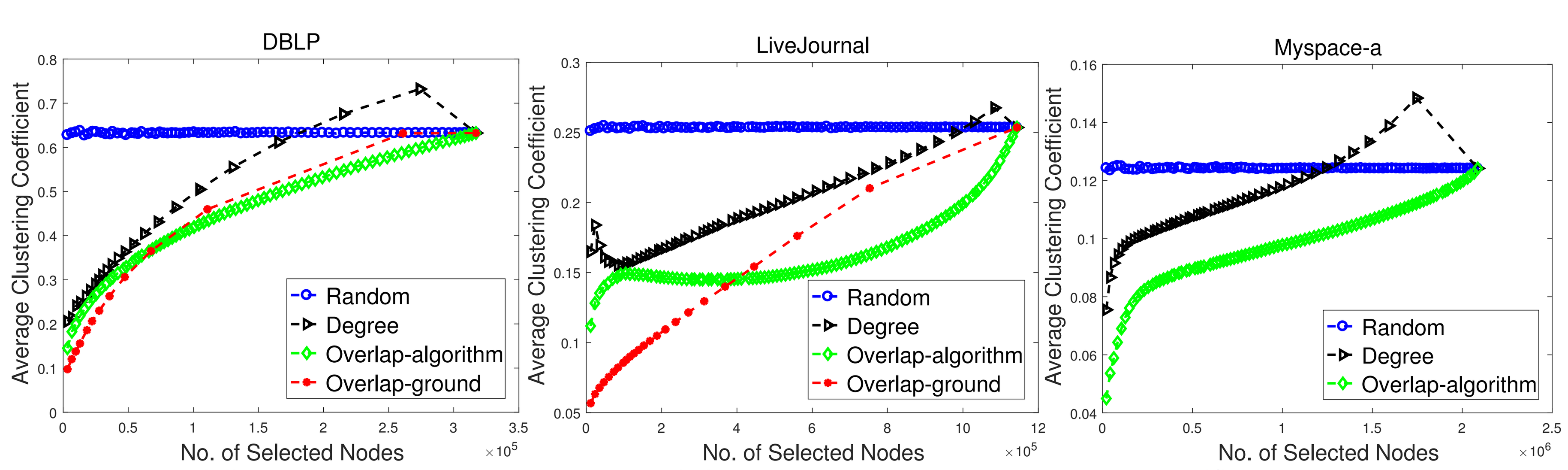
	DBLP	LiveJournal
No. of communities	13,477	662,859
No. of overlapped nodes (%)	110,806 (35%)	752,537 (65%)
No. of overlapped edges (%)	356,801 (34%)	4,724,058 (28%)

Overlapped Nodes and Structural Holes in a Network

- Structural hole: an empty space of a network between two sets of nodes that do not closely interact with each other.
  - A set of nodes that have multiple local bridges.
  - Adjacent to many local bridges $\longrightarrow$ a low clustering coefficient.

(Clustering coefficient of v_i: the probability that two randomly selected neighbors of v_i are directly connected.)

- Clustering coefficients of highly overlapped nodes
  - As the overlap degree (i.e., $|\mathcal{S}_i|$) increases, the average clustering coefficient decreases.
  - High-overlap nodes tend to have low clustering coefficients – even lower clustering coefficients than high-degree nodes.
  - Highly-overlapped nodes play as structural holes in a network.



(d) DBLP (e) LiveJournal (f) Myspace-a

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

- Sort the nodes according to their overlap degrees in descending order.
- t_p: the overlap degree of the $\lceil pn \rceil$-th node ($0 \leq p \leq 1$), n: the total # of nodes.
- Select the nodes whose overlap degrees are greater than or equal to t_p, and compute their average clustering coefficient.
- x-axis: $\lceil pn \rceil$, y-axis: the average clustering coefficient.

New Links in Community Overlap

- Social networks keep changing over time, e.g., new links are formed.
- Patterns of the link formations in the overlapped regions
  - Real-world datasets with the ground-truth new links.
  - New edges are formed within communities.
  - New edges are formed in the overlapped regions.
  - New edges include highly overlapped edges.

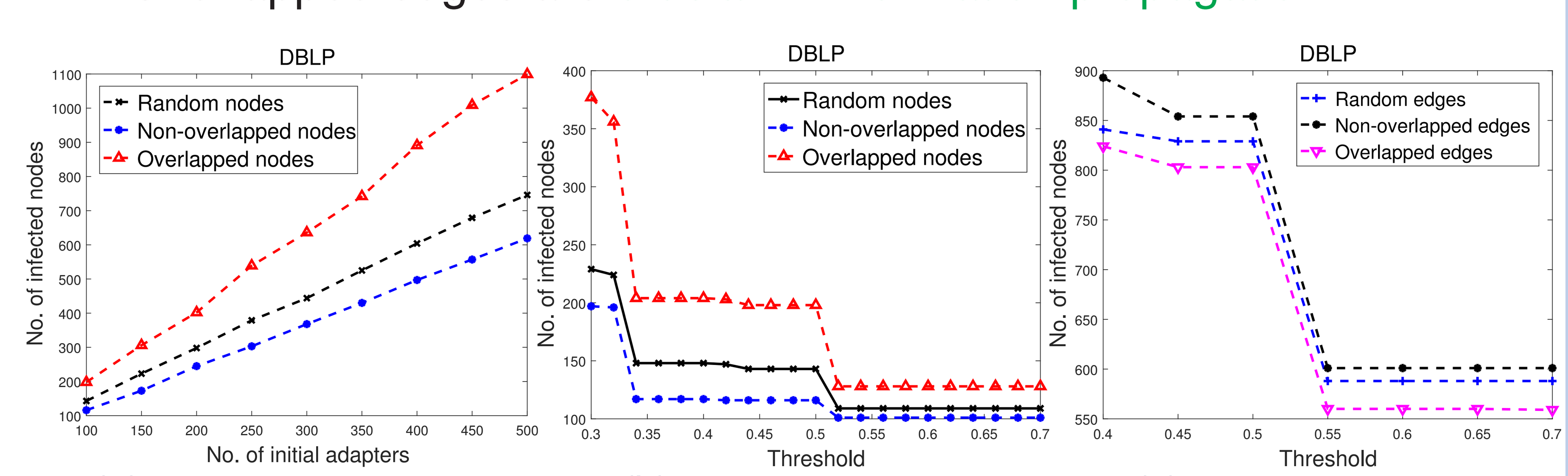
Table: Classification of the edges according to the number of common communities of the endpoints of the edges.

	Flickr-b		LiveJournal-b	
	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%)
$\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert = 1$	64,112 (16.19%)	103,164 (26.06%)	6,290 (0.97%)	99,433 (15.30%)
$\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert \geq 2$	257,910 (65.15%)	68,721 (17.36%)	634,679 (97.66%)	147,644 (22.72%)
mean($\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert$)	4.77	0.68	20.00	1.23
median($\lvert\mathcal{S}_i \cap \mathcal{S}_j\rvert$)	3	0	15	0

- $\mathcal{Q}$: the ground-truth new links, $\mathcal{R}$: randomly generated links
- Given an edge $e = \{v_i, v_j\}$, we classify the edge into three categories: (i) a between-community edge, (ii) a non-overlapped within community edge, (iii) an overlapped edge.

Information Diffusion through Overlapped Nodes and Edges

- Information diffusion: model the way how information is propagated.
- A networked coordination game
  - Each node has a choice between two possible behaviors A and B.
  - 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.
- Roles of the overlapped nodes
  - The number of infected nodes is maximized when we select the initial nodes among the overlapped nodes.
  - Whether a node is an overlapped node or not is an important factor to determine the success of information spreading.
  - The overlapped nodes effectively spread the information.
- Roles of the overlapped edges
  - Information is not spread well when the overlapped edges are removed.
  - Overlapped edges are crucial in information propagation.



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

Figure: (a)&(b): Information diffusion with different initial nodes. (c) Information diffusion with differently removed edges.

- We choose the initial node set in three different ways: (i) random nodes, (ii) non-overlapped nodes, and (iii) overlapped nodes.
- We remove edges in the network in three ways: (i) random edges, (ii) non-overlapped edges, and (iii) overlapped edges.

Conclusions & Future Work

- High-overlap nodes have low clustering coefficients–they bridge different communities.
- When networks evolve over time, the new links tend to be formed within overlapped regions.
- Overlapped nodes and overlapped edges play a critical role in spreading information throughout the network.
- Useful intuition and insight for many practical applications including link prediction and information propagation models.

Joyce Jiyoung Whang. An Empirical Study of Community Overlap: Ground-truth, Algorithmic Solutions, and Implications. In CIKM, 2017.