Mining Structural Hole Spanners in Social Networks

Jie Tang

Department of Computer Science and Technology
Tsinghua University

Social Networks

- >**1000 million** users
- The **3rd** largest "Country" in the world
- More visitors than Google
- >**800 million** users
- 2013, **560 million** users, 40% yearly increase

- 2009, **2 billion** tweets per quarter
- 2010, **4 billion** tweets per quarter
- 2011,**25 billion** tweets per quarter
- More than **6 billion** images
- Pinterest, with a traffic higher than Twitter and Google

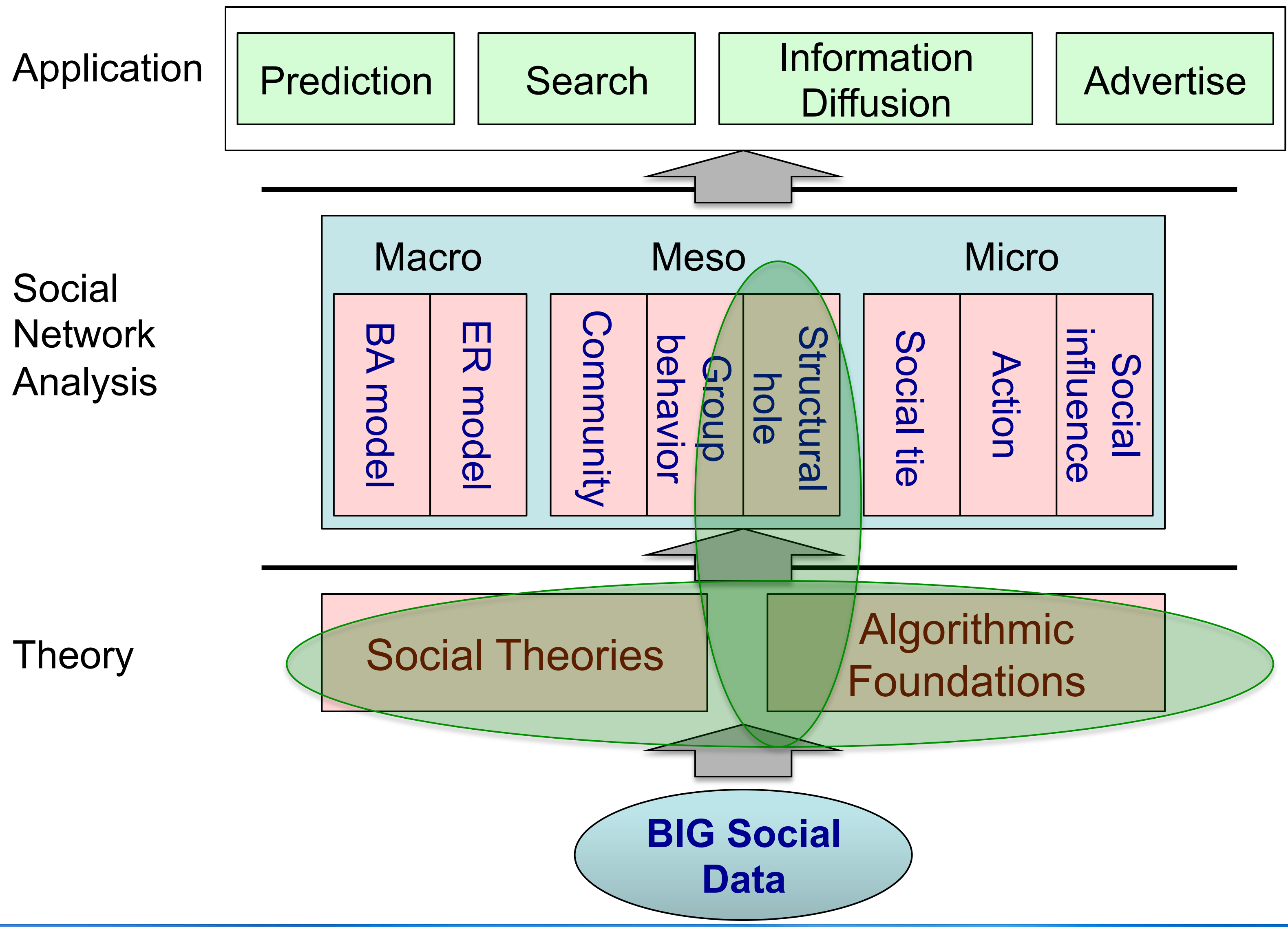
A Trillion Dollar Opportunity

Social networks already become a bridge to connect our daily **physical** life and the **virtual** web space

On2Off [1]

[1] Online to Offline is trillion dollar business
http://techcrunch.com/2010/08/07/why-online2offline-commerce-is-a-trillion-dollar-opportunity/

Core Research in Social Network

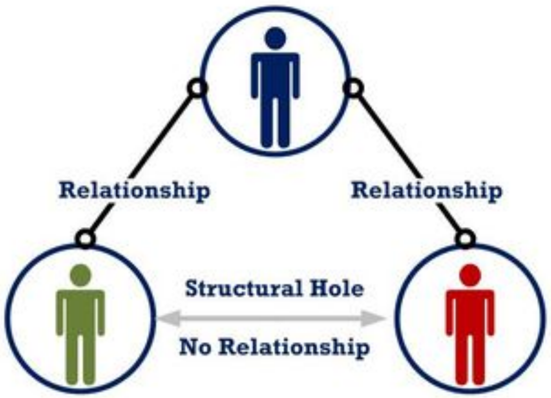


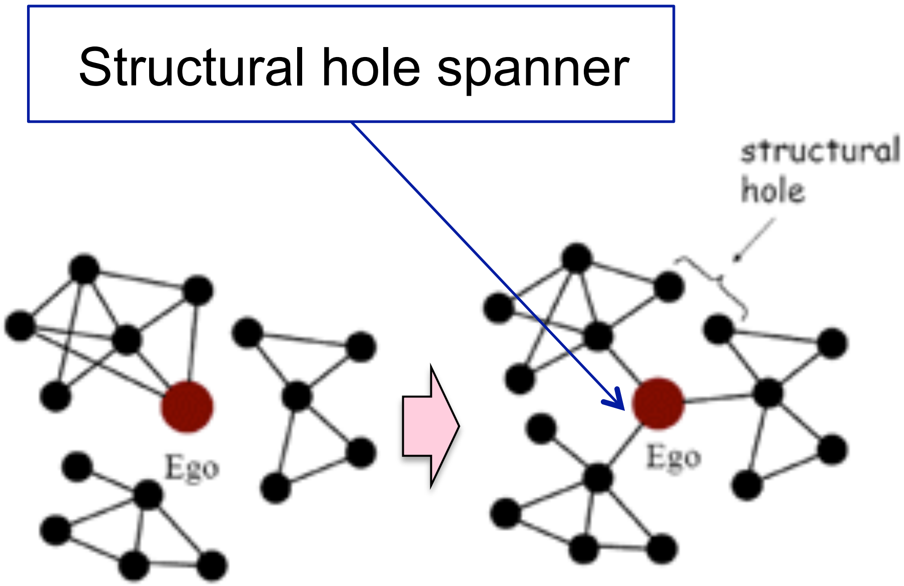
Today, let us start with the notion of "structural hole"…

What is "Structural Hole"?

- Structural hole: When two separate clusters possess non-redundant information, there is said to be a structural hole between them.[1]

Structural hole spanner

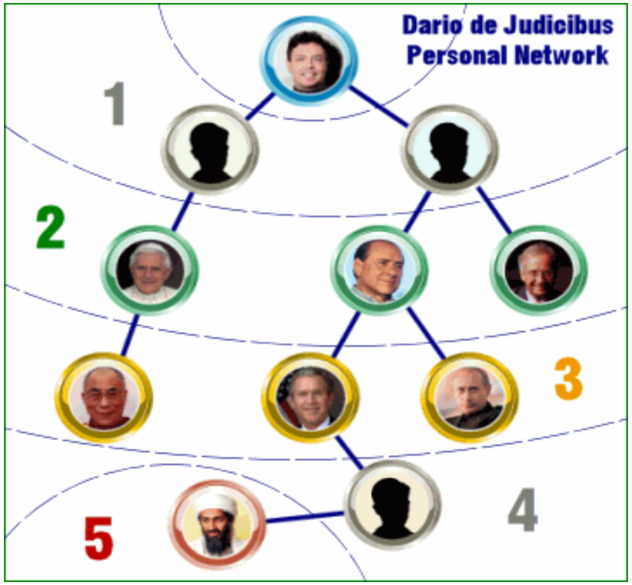




[1] R. S. Burt. Structural Holes: The Social Structure of Competition. Harvard University Press, 1992.

Few People Connect the World

Six degree of separation[1]



In that famous experiment…

- Half the arrived letters passed through the same three people.
- It's not about how we are connected with each other. It's about how we are linked to the world through few "gatekeepers"[2].
- How could the letter from a painter in Nebraska been received by a stockbroker in Boston?

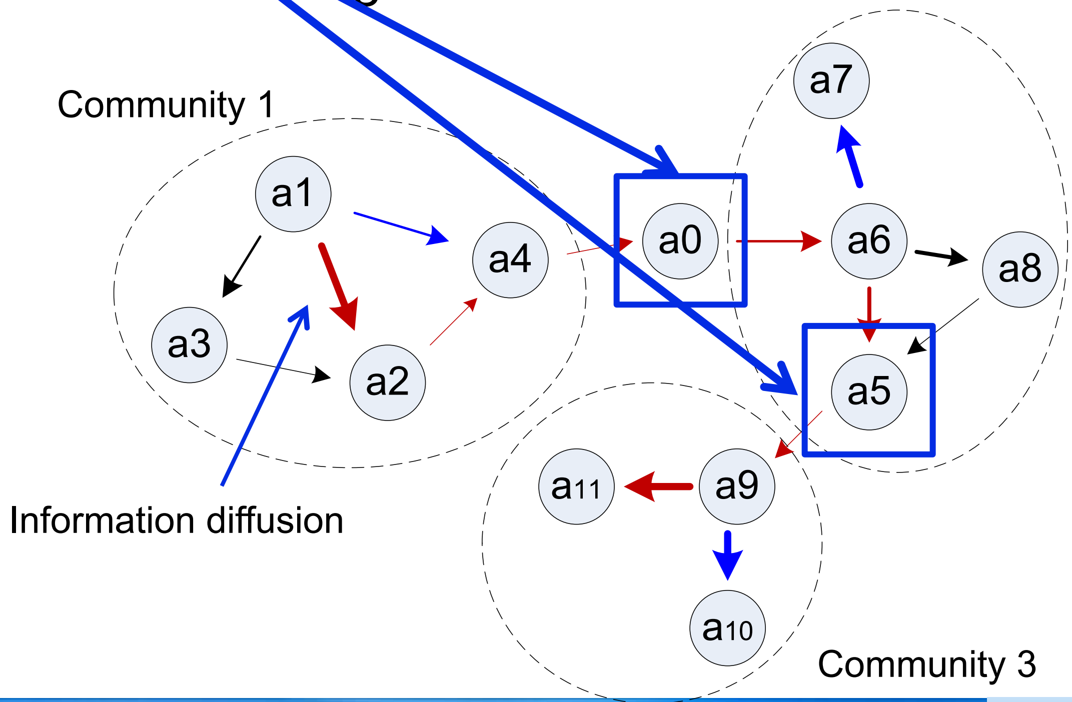
[1] S. Milgram. The Small World Problem. Psychology Today, 1967, Vol. 2, 60–67

[2] M. Gladwell. The Tipping Point: How Little Things Can Make A Big Difference. 2006.

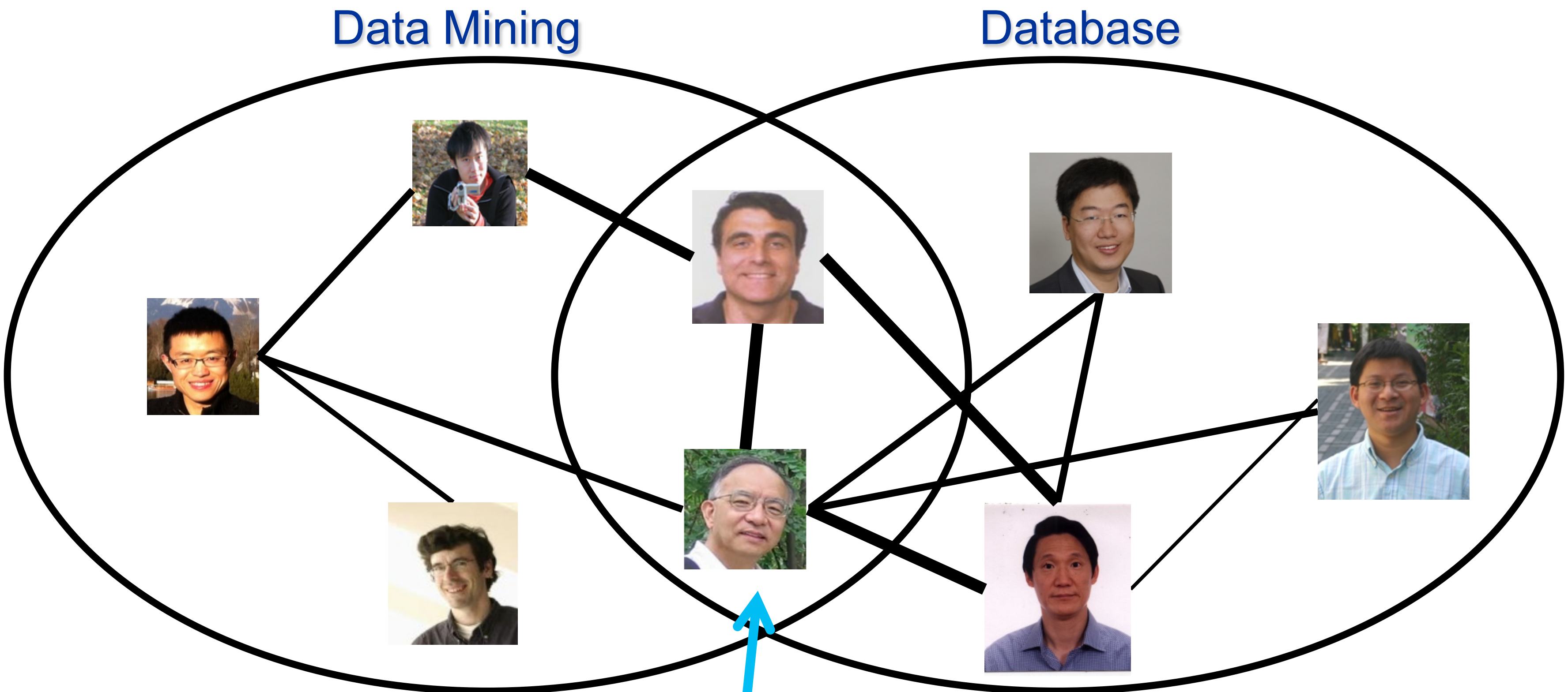
Structural hole spanners control information diffusion…

- The theory of Structural Hole [Burt92]:
  - "Holes" exists between communities that are otherwise **disconnected**.
- Structural hole spanners
  - Individuals would **benefit** from filling the "holes".

On Twitter, **Top 1%** twitter users control **25%** retweeting flow between communities.



Examples of DBLP & Challenges



Challenge 1 : Stru
spanner vs Opini

82 overlapped PC members of **SIGMOD/ICDT/VLDB** and **SIGKDD/ICDM** during years 2007 – 2009.

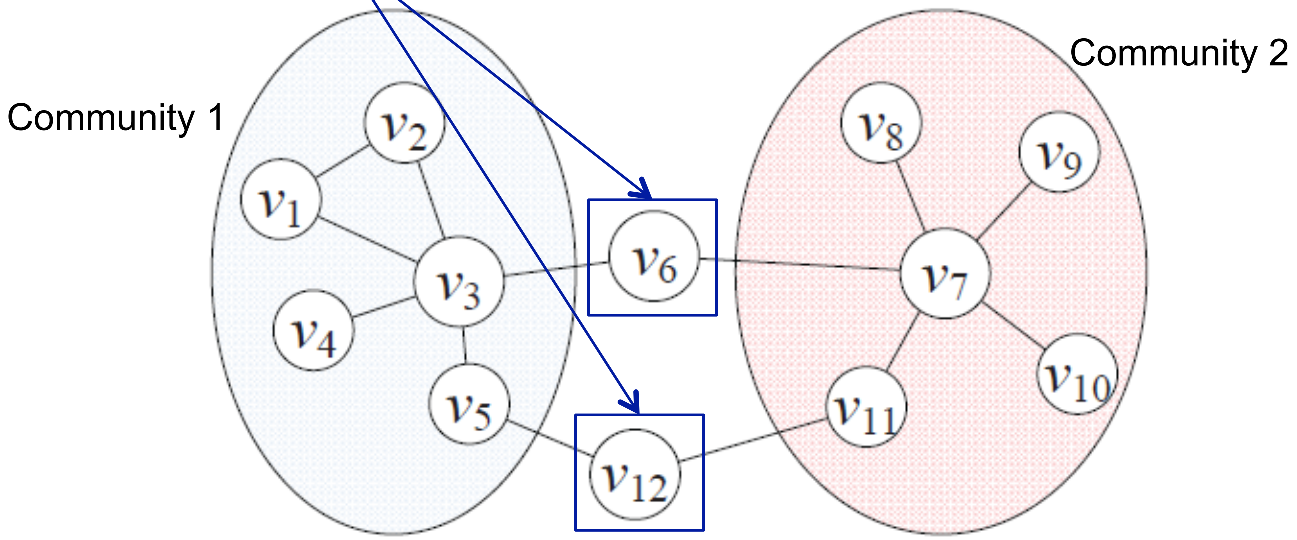
: Who control
tion diffusion?

Mining Top-k Structural Hole Spanners

[1] T. Lou and J. Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In **WWW'13**. pp. 837-848.

Problem Definition

Which node is the best structural hole spanner?



Well, mining top-k structural hole spanners is more complex…

Problem definition

- INPUT :
  - A social network, $G = (V, E)$ and L communities $C = (C_1, C_2, \ldots, C_L)$
- Identifying top-k structural hole spanners.

$$\max Q(V_{SH}, C), \text{ with } |V_{SH}| = k$$

Utility function Q(V*, C) : measure V*'s degree to span structural holes.

V_{SH} : Top-k structural holes spanners as a subset of k nodes

Data

	#User	#Relationship	#Messages
Coauthor	815,946	2,792,833	1,572,277 papers
Twitter	112,044	468,238	2,409,768 tweets
Inventor	2,445,351	5,841,940	3,880,211 patents

- In **Coauthor**, we try to understand how authors bridge different research fields (e.g., DM, DB, DP, NC, GV);
- In **Twitter**, we try to examine how structural hole spanners control the information diffusion process;
- In **Inventor**, we study how technologies spread across different companies via inventors who span structural holes.

Our first questions

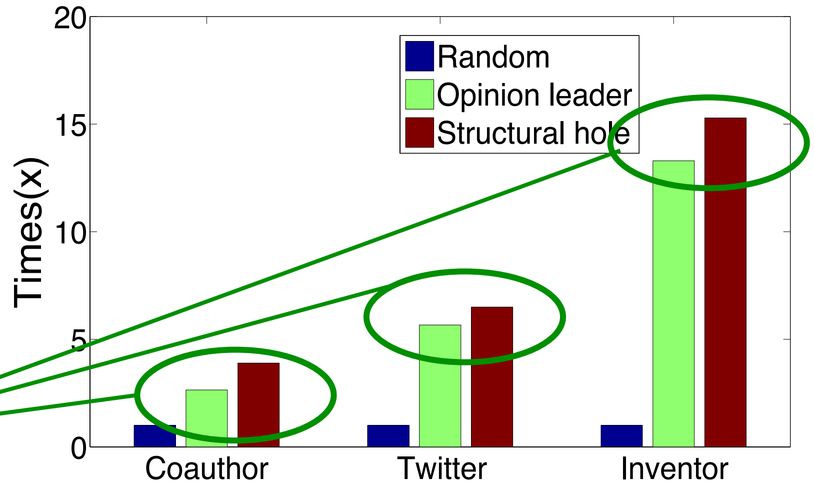
- Observable analysis
  - How likely would **structural hole spanners** connect with "**opinion leaders**" ?
  - How likely would **structural hole spanners** influence the "**information diffusion**"?

Structural hole spanners vs Opinion leaders

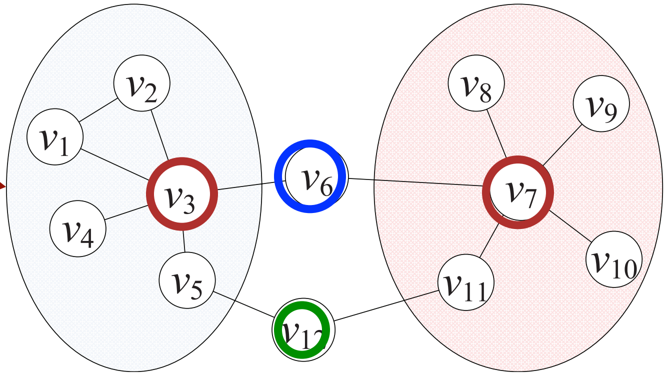
Structural hole vs. Opinion leader vs. Random

Result: Structural hole spanners are more likely to connect important nodes

+15% - 50%

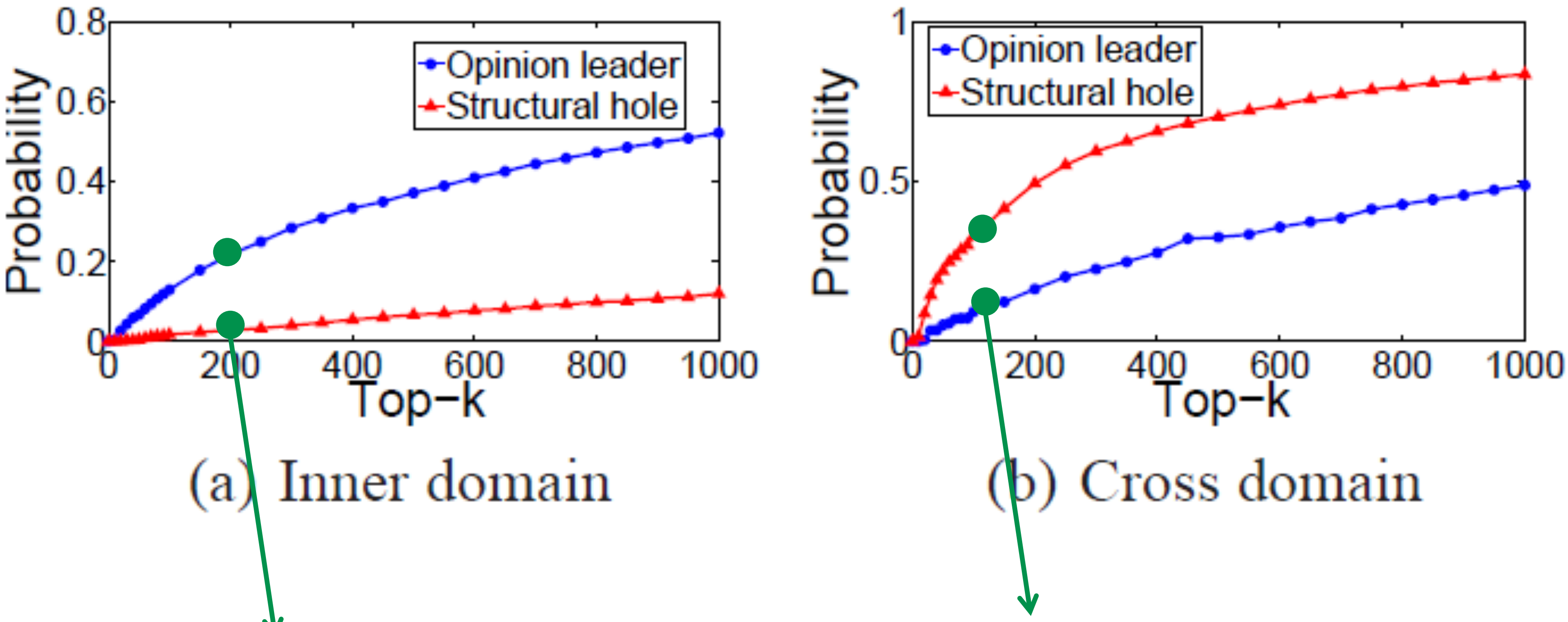


The two-step information flow theory[1] suggests structural hole spanners are connected with many "opinion leaders"



[1] E. Katz. The two-step flow of communication: an up-to-date report of an hypothesis. In Enis and Cox(eds.), Marketing Classics, pages 175–193, 1973.

Structural hole spanners control the information diffusion

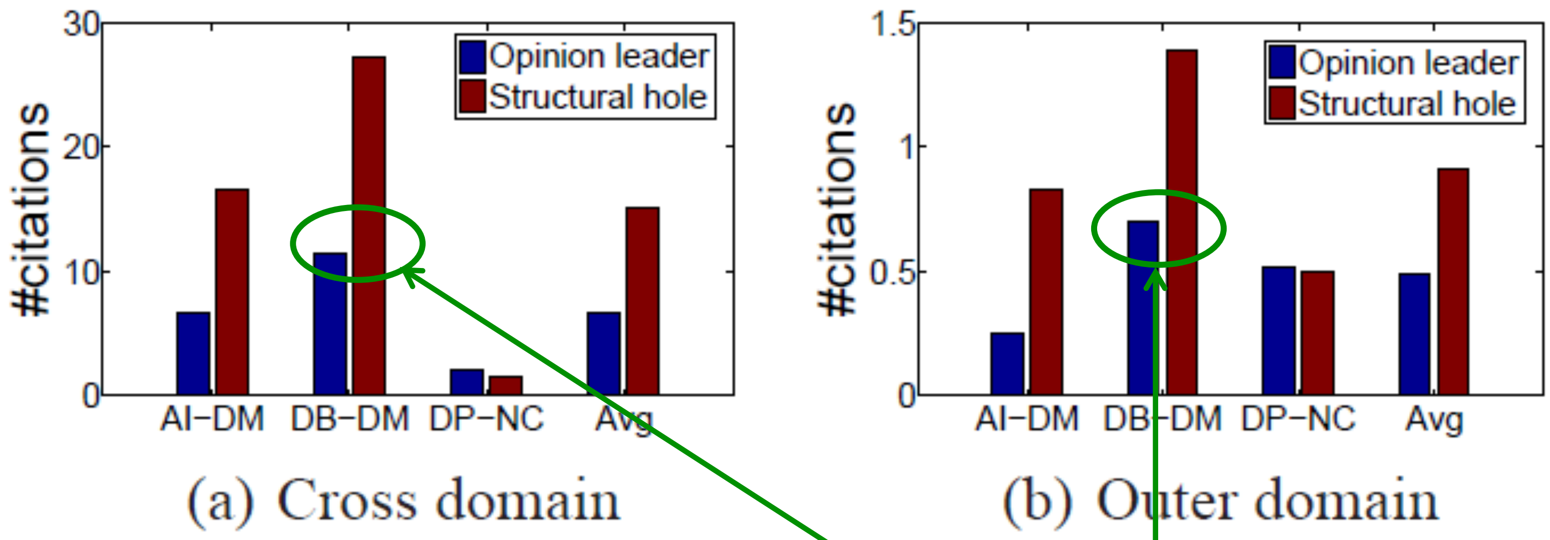


Opinion leaders 5 times higher

Structural hole spanners 3 times higher

Results: **Opinion leaders** controls information flows within communities, while **Structural hole spanners** dominate information spread across **communities**.

Structural hole spanners influence the information diffusion



In the **Coauthor** network :

Structural hole spanners almost **double** opinion leaders on number of **cross** domain (and **outer** domain) citations.

Intuitions

- Structural hole spanners are more likely to connect important nodes in different communities.

  → Model 1 : HIS

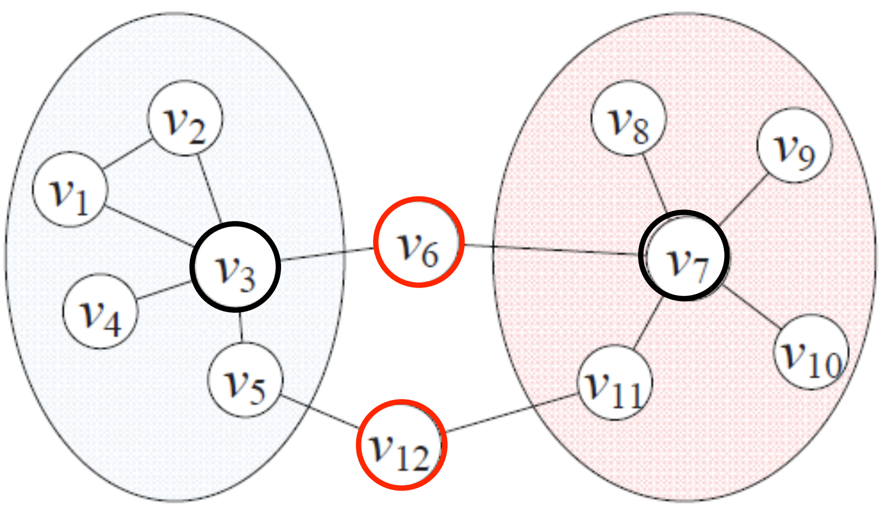
- Structural hole spanners control the information diffusion between communities.

  → Model 2 : MaxD

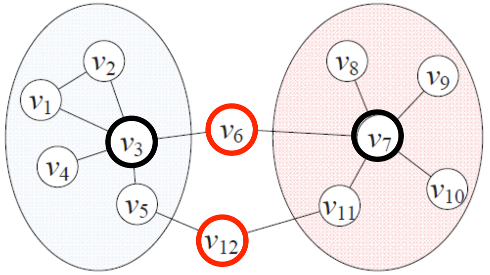
Models, Algorithms, and Theoretical Analysis

Model One : HIS

- Structural hole spanners are more likely to connect important nodes in different communities.
  - If a user is connected with many opinion leaders in different communities, more likely to span structural holes.
  - If a user is connected with structural hole spanners, more likely to act as an opinion leader.



Model One : HIS



- Structural hole spanners are more likely to connect important nodes in different communities.
  - If a user is connected with many opinion leaders in different communities, more likely to span structural holes.
  - If a user is connected with structural hole spanners, more likely to act as an opinion leader.
- Model
  - $\boxed{I(v, C_i)} = max\ \{\ I(v, C_i), \boxed{\alpha_i} I(u, C_i) + \boxed{\beta_S} H(u, S)\ \}$
  - $\boxed{H(v, S)} = min\ \{\ I(\mathrm{v}, \mathrm{C_i})\ \}$

I(v, C_i) : importance of *v* in community C_i.
H(v, S) : likelihood of *v* spanning structural holes across *S* (subset of communities).

α and β are two parameters

Algorithm for HIS

Input: $G = (V, E)$, parameters α_i, β_S, and convergence threshold ϵ
Output: Importance I and structural hole score H

Initialize $I(v, C_i)$ according to $\boxed{\text{Eq. 4 ;}}$
repeat
    foreach $v \in V$ **do**
        foreach $C_i \in \mathbf{C}$ **do**
            $P(v, C_i) =$
            $\max_{S \subseteq \mathbf{C} \wedge C_i \in S}\{\alpha_i I(v, C_i) + \beta_S H(v, S)\}$;
        end
    end
    foreach $v \in V$ **do**
        foreach $C_i \in \mathbf{C}$ **do**
            $I'(v, C_i) = \max\{I(v, C_i), \max_{e_{uv} \in E} P(u, C_i)\}$;
        end
        foreach $S \subseteq \mathbf{C}$ **do**
            $H'(v, S) = \min_{C_i \in S} I'(v, C_i)$;
        end
    end
    Check the ϵ-convergence condition by

$$\max_{v \in V, C_i \in \mathbf{C}} |I'(v, C_i) - I(v, C_i)| \boxed{\leq \epsilon}$$

    Update $I = I'$ and $H = H'$;
until *Convergence*;

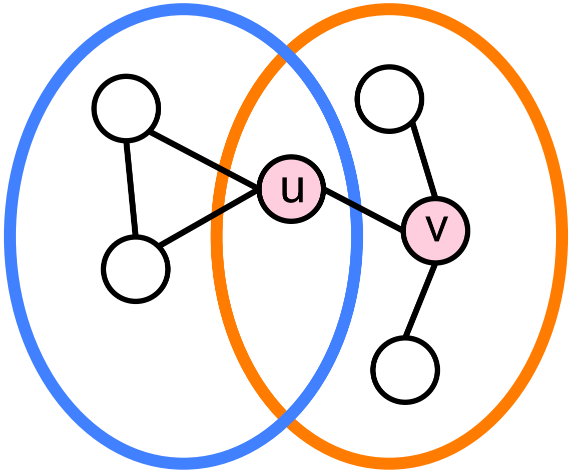
$$I(v, C_i) = \underline{r(v)}, \quad v \in C_i$$
$$I(v, C_i) = 0, \quad v \notin C_i$$

By PageRank
or HITS

Parameter to control the convergence

Theoretical Analysis—Existence

- Given α_i and β_S, solution exists ($I(v, C_i)$, $H(v, S) \leq 1$) for any graph, if and only if, $\alpha_\mathrm{i} + \beta_\mathrm{S} \leq 1$.
  - For the *only if* direction
    - $Suppose\ \alpha_i + \beta_S > 1$, $S = \{C_{\mathrm{blue}}, C_{\mathrm{yellow}}\}$
    - $r(u) = r(v) = 1;$
    - $I(\mathrm{u}, \mathrm{C}_{\mathrm{blue}}) = I(\mathrm{u}, \mathrm{C}_{\mathrm{yellow}}) = 1;$
    - $H(\mathrm{u}, \mathrm{S}) = min\ \{ I(\mathrm{u}, \mathrm{C}_{\mathrm{blue}}), I(\mathrm{u}, \mathrm{C}_{\mathrm{yellow}})\} = 1;$
    - $I(v, C_{\mathrm{yellow}}) \geq \alpha_i\, I(u, C_i) + \beta_S\, H(u, S) = \alpha_i + \beta_S > 1$



$I(v, C_i) = \max \{ I(v, C_i), \alpha_i\, I(u, C_i) + \beta_S\, H(u, S) \}$
$H(v, S) = \min \{ I(v, C_i) \}$

Theoretical Analysis—Existence

- Given α_i and β_S, solution exists ($I(v, C_i)$, $H(v, S) \leq 1$) for any graph, if and only if, $\alpha_i + \beta_S \leq 1$.
  - For the *if* direction
    - *If* $\alpha_i + \beta_S \leq 1$, we use induction to prove $I(v, C_i) \leq 1$;
    - Obviously $I^{(0)}(v, C_i) \leq r(v) \leq 1$;
    - Suppose after the k-th iteration, we have $I^{(k)}(v, C_i) \leq 1$;
    - Hence, in the (k + 1)-th iteration, $I^{(k+1)}(v, C_i) \leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k)}(u, S) \leq (\alpha_i + \beta_S) I^{(k)}(u, C_i) \leq 1$.

$I(v, C_i) = \max \{ I(v, C_i), \alpha_i I(u, C_i) + \beta_S H(u, S) \}$
$H(v, S) = \min \{ I(v, C_i) \}$

Theoretical Analysis—Convergence

- Denote $\gamma = \alpha_i + \beta_S \leq 1$, we have

$$|I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$$

  – When $k = 0$, *we have* $I^{(1)}(v, C_i) \leq 1$, *thus*

$$|I^{(1)}(v, C_i) - I^{(0)}(v, C_i)| \leq 1$$

  – Assume after *k-th iteration, we have*

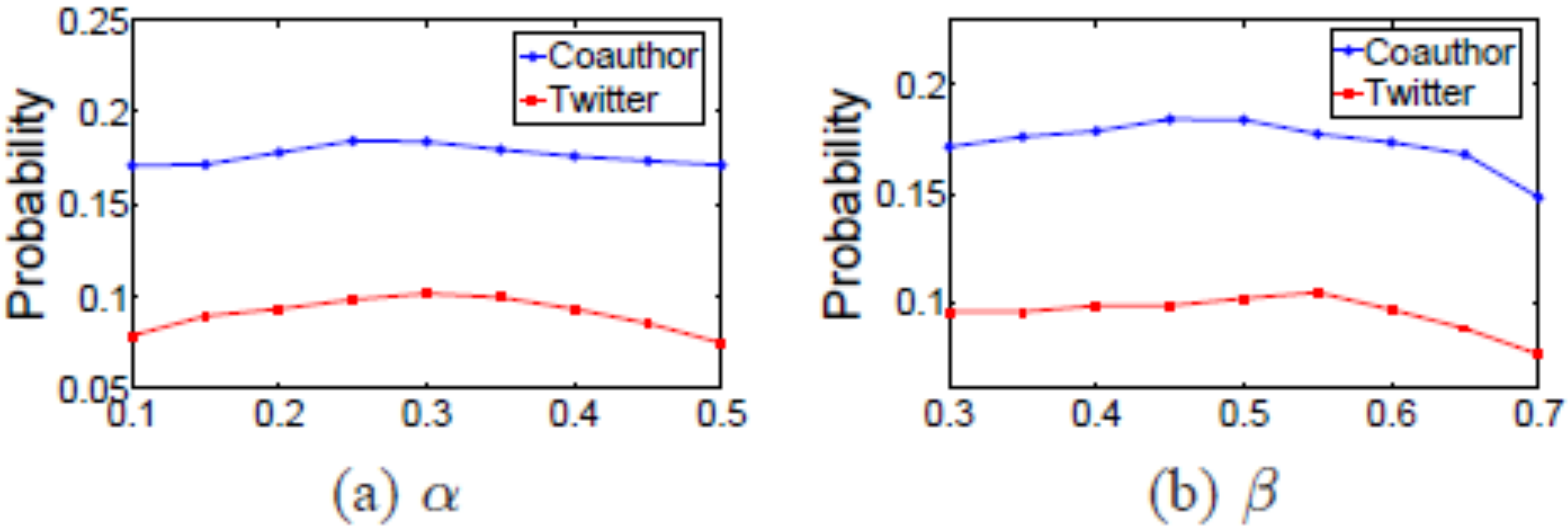
$$|I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$$

  – After $(k+1)$*-th iteration, we have*

$$\begin{aligned} I^{(k+2)}(v, C_i) &= \alpha_i I^{(k+1)}(u, C_i) + \beta_S H^{(k+1)}(u, S) \\ &\leq \alpha_i[I^{(k)}(u, C_i) + \gamma^k] + \beta_S[H^{(k+1)}(u, S) + \gamma^k] \\ &\leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k+1)}(u, S) + \gamma^{k+1} \\ &\leq I^{(k+1)}(u, C_i) + \gamma^{k+1} \end{aligned}$$

Convergence Analysis

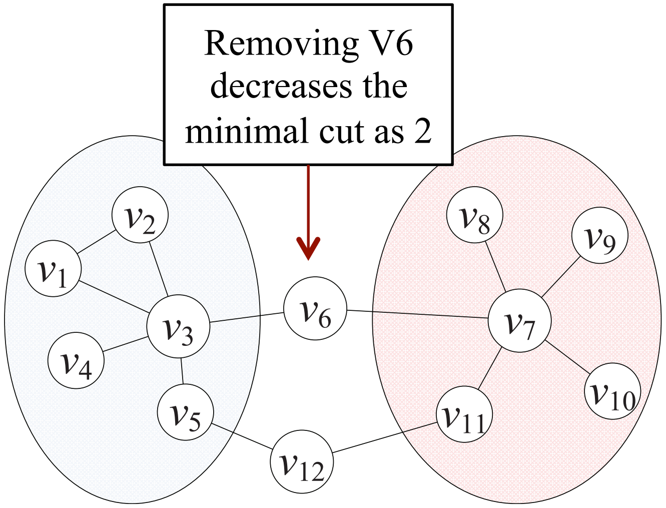
- Parameter analysis.
  - The performance is insensitive to the different parameter settings.



Model Two: MaxD

- The minimal cut D of a set communities C is the minimal number of edges to separate nodes in different communities.

- The structural hole spanner detection problem can be cast as finding top-k nodes such that after removing these nodes, the decrease of the minimal cut will be maximized.



Two communities with the minimal cut as 4

Model Two: MaxD

- Structural holes spanners play an important role in **information diffusion**

$$Q(V_{SH}, C) = \boxed{\mathrm{MC}}(G, C) - \boxed{\mathrm{MC}}(G \setminus V_{SH}, C)$$

MC(G, C) = the minimal cut of communities C in G.

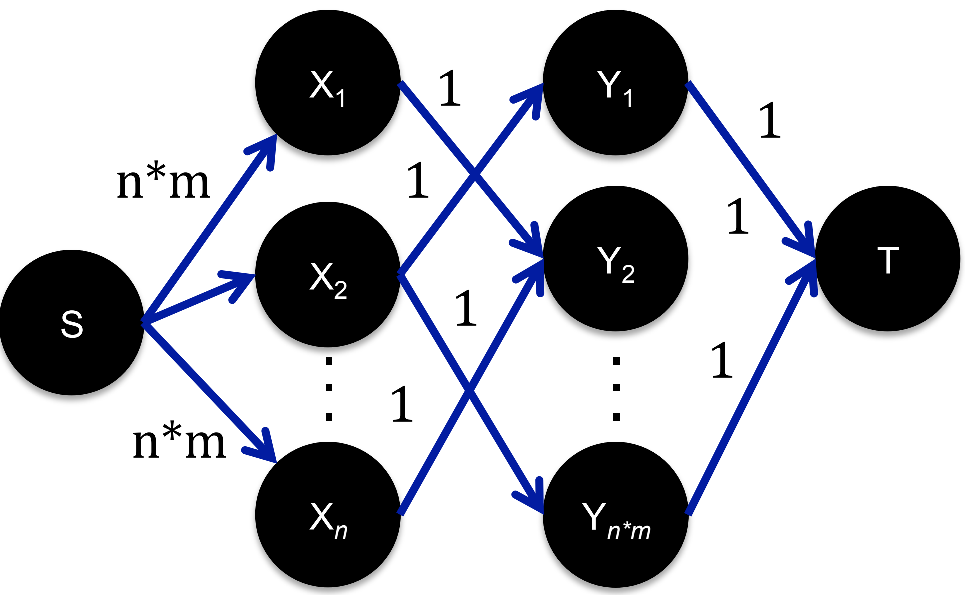
Hardness Analysis

$$Q(V_{SH}, C) = \mathrm{MC}\ (G, C) - \mathrm{MC}\ (G \setminus V_{SH}, C)$$

- Hardness analysis
  - If $|V_{SH}| = 2$, the problem can be viewed as **minimal node-cut problem**
  - We already have NP-Hardness proof for **minimal node-cut problem**, but the graph is exponentially weighted.
  - Proof NP-Hardness in an un-weighted (polybounded -weighted) graph, by reduction from **k-DENSEST-SUBGRAPH** problem.

Hardness Analysis

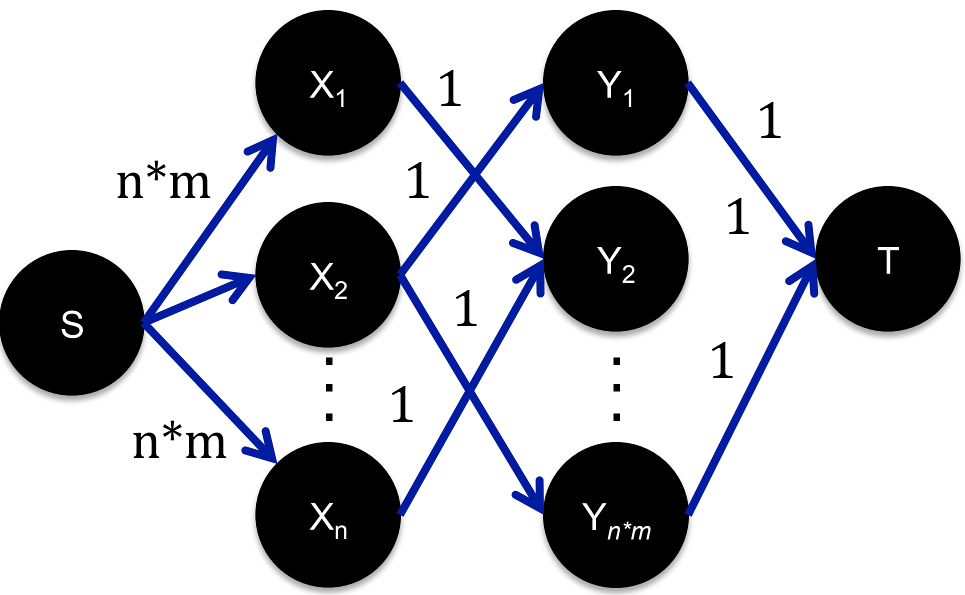
- Let us reduce the problem to an instance of the k-DENSEST SUBGRAPH problem



- Given an instance {G'=<V, E>, k, d} of the k-DENSEST SUBGRAPH problem, n=|V|, m=|E|;
- Build a graph G with a source node S and target node T;
- Build n nodes connecting with S with capacity $n*m$;
- Build n nodes for each edge in G', connect each of them to T with capacity 1;

Hardness Analysis (cont.)

- Build a link from x_i to y_j with capacity 1 if the x_i in G' appears on the j-th edge;
- $MC(G)=n*m$;



- The instance is satisfiable, if and only if there exists a subset

$$|V_{SH}|=k$$

such that

$$MC(G\backslash V_{SH}) <= n(m-d)$$

Proof: NP-hardness (cont.)

- For the *only if* direction
  - Suppose we have a sub-graph consists of k nodes {x'} and at least d edges;
  - We can choose V_{SH}={x};
  - For the k-th edge y in G', if y exists in the sub-graph, two nodes appearing on y are removed in G;
  - Thus y cannot be reached and we lost n flows for y;
  - Hence, we have MC($G \setminus V_{SH}$) <= n*(m-d).

Proof: NP-hardness (cont.)

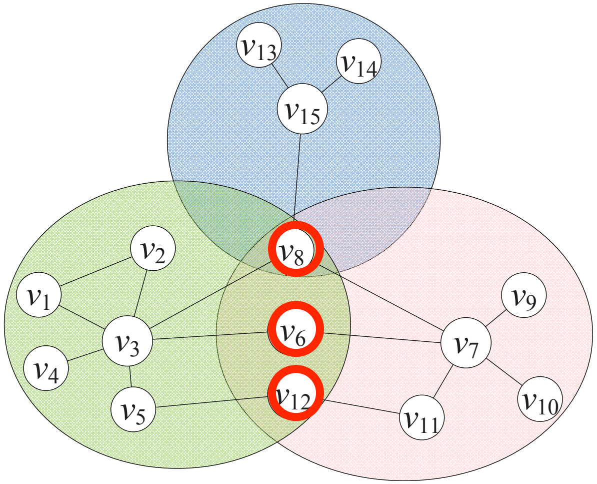
- For the *if* direction
  - If there exists a k-subset V_{SH} such that MC(G\\V_{SH}) <= n*(m-d);
  - Denote V_{SH}'=V_{SH}^{x}, the size of V_{SH}' is at most k, and MC(G\\V_{SH}') <= n*(m-d);
  - Let the node set of the sub-graph be V_{SH}', thus there are at least *d* edges in that sub-graph.

Approximation Algorithm

- Two approximation algorithms:
  - Greedy: in each iteration, select a node which will result in a max-decrease of $Q(.)$ when removed it from the network.
  - Network-flow: for any possible partitions E_S and E_T, we call a network-flow algorithm to compute the minimal cut.

An example: finding top 3 structural holes

Step 1: select V8 and decrease the minimal cut from 7 to 4
Step 2: select V6 and decrease the minimal cut from 4 to 2
Step 3: select V12 and decrease the minimal cut from 2 to 0



Approximation Algorithm

Greedy : In each round, choose the node which results in the max-decrease of *Q*.

Input: $G = (V, E), k, l, \mathbf{C} = \{C_i\})$
Output: Top-k structural hole nodes V_{SH}

Initialize $V_{SH} = \emptyset$;
while $|V_{SH}| < k$ **do**
  Initialize $f(v) = 0$, for each $v \in V$;
  foreach *non empty* $S \subset \{1, \cdots, l\}$ **do**
    $E_S = \cup_{i \in S} C_i$ and $E_T = \cup_{i \notin S} C_i$;
    Compute the maximal flow with source E_S and sink E_T on the induced graph $G \setminus V_{SH}$;
    foreach $v \in V$ **do**
      Add $f(v)$ by the flow though node v ;
    end
  end
  Choose $O(k)$ nodes with the largest f as candidates D;
  Compute $p^* = \arg\min_{p \in D} MC(G \setminus (V_{SH} \bigcup \{p\}), \mathbf{C})$;
  Update $V_{SH} = V_{SH} \bigcup \{p^*\}$
end

Step 1: Consider top O(k) nodes with maximal sum of flows through them as candidates.

Step 2: Compute MC(*, *) by trying all possible partitions.

Complexity: $O(2^{2l} T_2(n))$; $T_2(n)$—the complexity for computing min-cut
Approximation ratio: $O(\log l)$

Results

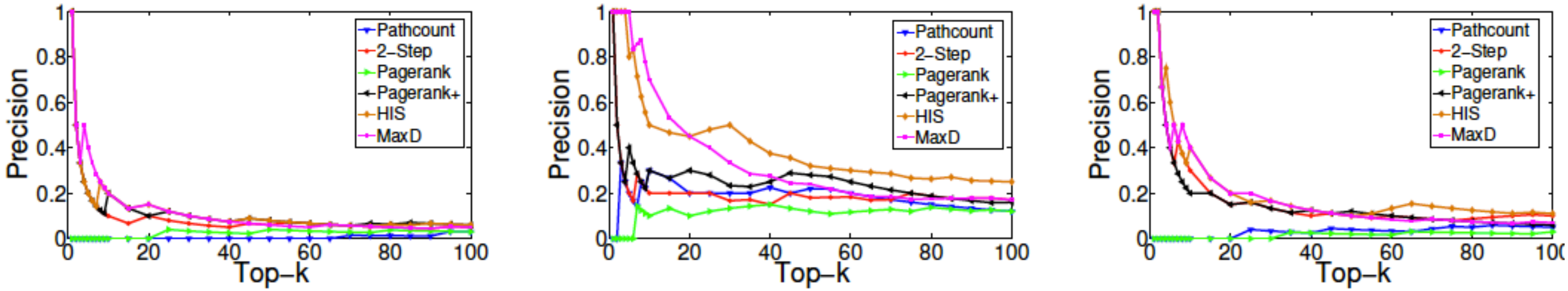
Experiment

	#User	#Relationship	#Messages
Coauthor	815,946	2,792,833	1,572,277 papers
Twitter	112,044	468,238	2,409,768 tweets
Inventor	2,445,351	5,841,940	3,880,211 patents

- Evaluation metrics
  - Accuracy (Overlapped PC members in the Coauthor network)
  - Information diffusion on Coauthor and Twitter.
- Baselines
  - Pathcount: #shortest path a node lies on
  - 2-step connectivity: #pairs of disconnected neighbors
  - Pagerank and PageRank+: high PR in more than one communities

Experiments

- Accuracy evaluation on Coauthor network



(a) AI-DM (b) DB-DM (c) DP-NC

- Predict overlapped PC members on the Coauthor network.
  - +20 – 40% on precision of AI-DM, DB-DM and DP-NC
- What happened to AI-DM?

Experiment results (accuracy)

- What happened to AI-DB?
  - Only 4 overlapped PC members on AI and DB during 2007 – 2009, but 40 now.
  - Our conjecture : **dynamic of structural holes**.

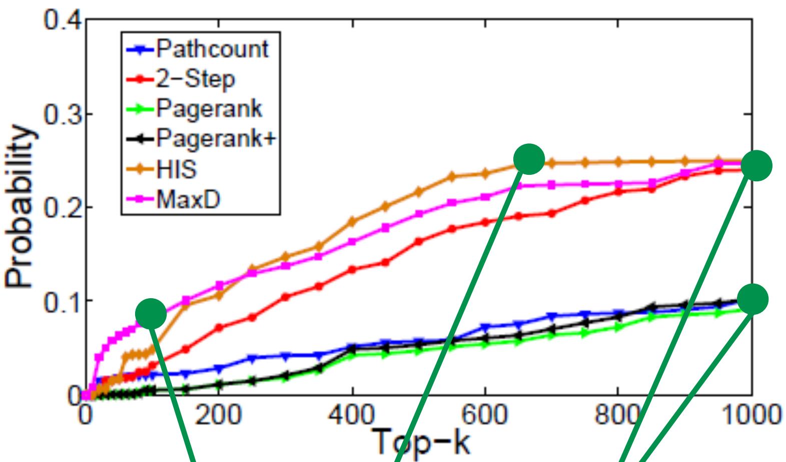
Structural holes spanners of **AI** and **DB** form the **new area DM**.

Similar pattern for
1) Collaborations between experts in AI and DB.
2) Influential of **DM** papers.

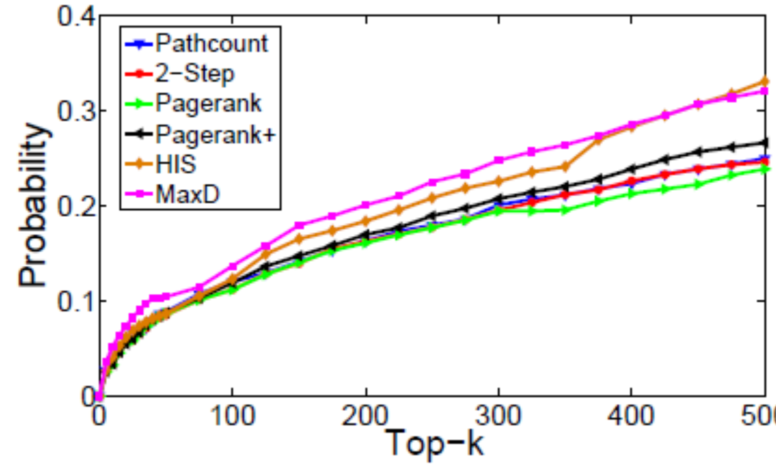
Significantly increase of coauthor links of AI and DB around year **1994**.

Most overlapped PC members on AI and DB are also PC of **SIGKDD**

Maximization of Information Spread



(a) Twitter



(b) Coauthor

Clear improvement. **(2.5 times)**

Top 0.2% - 10 %
Top 1% - 25 %

Improvement is limited, due to top a few authors dominate.

Improvement is statistically significant ($p << 0.01$)

Case study on the inventor network

- Most structural holes have more than one jobs.
- Mark * on inventors with highest PageRank scores.
  - HIS selects people with highest PageRank scores,
  - MaxD tends to select people how have been working on more jobs.

Inventor	HIS	MaxD	Title
E. Boyden		√	Professor (MIT Media Lab)
			Associate Professor (MIT McGovern Inst.)
			Group Leader (Synthetic Neurobiology)
A.A. Czarnik		√	Founder and Manager (Protia, LLC)
			Visiting Professor (University of Nevada)
			Co-Founder (Chief Scientific Officer)
A. Nishio		√	Director of Operations (WBI)
			Director of Department Responsible (IDA)
E. Nowak*	√		Senior vice President (Walt Disney)
			Secretary of Trustees (The New York Eye)
A. Rofougaran	√		Consultant (various wireless companies)
			Co-founder (Innovent System Corp.)
			Leader (RF-CMOS)
S. Yamazaki*	√		President and majority shareholder (SEL)

Efficiency

- Running time of different algorithms in three data sets

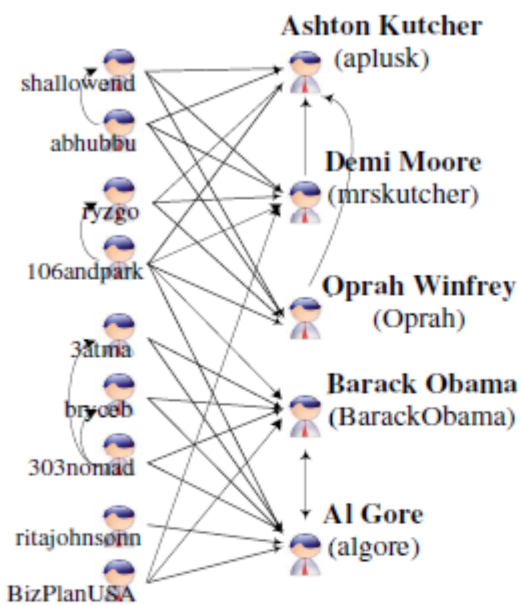
Data Set	Pathcount	2-Step	PageRank	HIS	MaxD
Coauthor	350.66s	4.71s	0.20s	0.60s	189.78m
Twitter	32.03m	12.09s	0.67s	3.87s	602.37m
Inventor	494.3 hr	98.96s	3.61s	26.11s	370.8hr

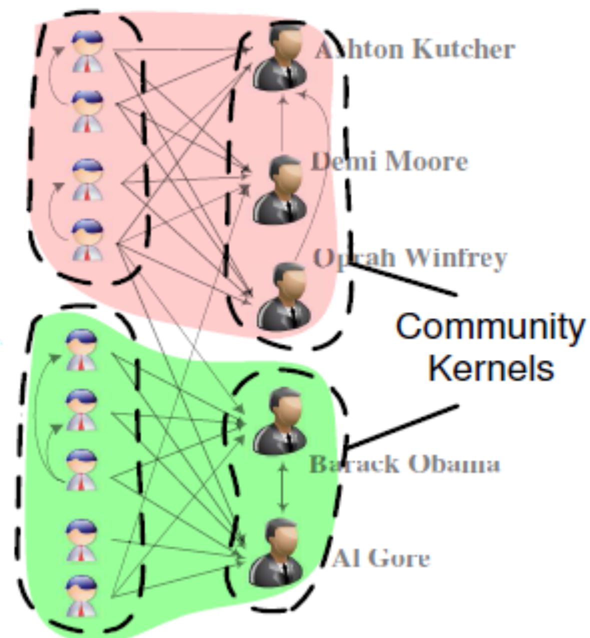
Inefficient!!

Applications

Detecting Kernel Communities

- Community kernel detection
  - GOAL : obtain the importance of each **node** within each **community** (as **kernel members**).
  - HOW : kernel members are **more** likely to connect structural hole spanners.

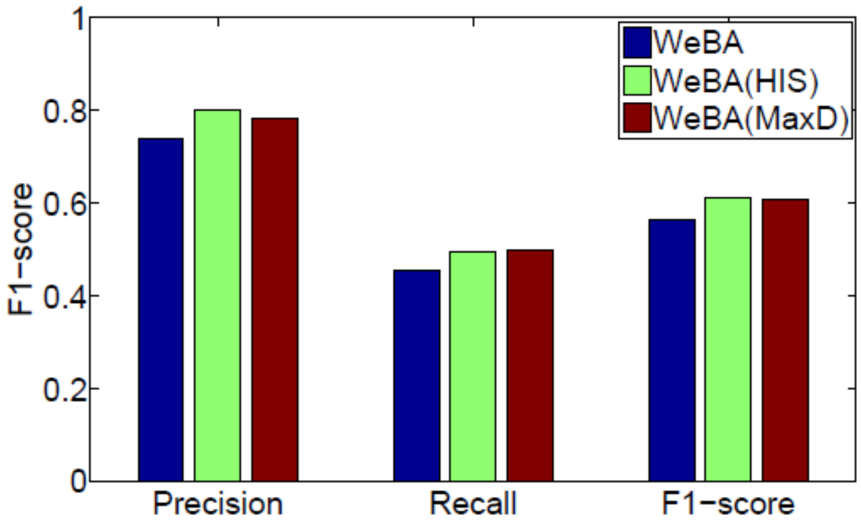




[1] L. Wang, T. Lou, J. Tang, and J. E. Hopcroft. Detecting Community Kernels in Large Social Networks. In **ICDM'11**. pp. 784-793.

Detecting Kernel Communities

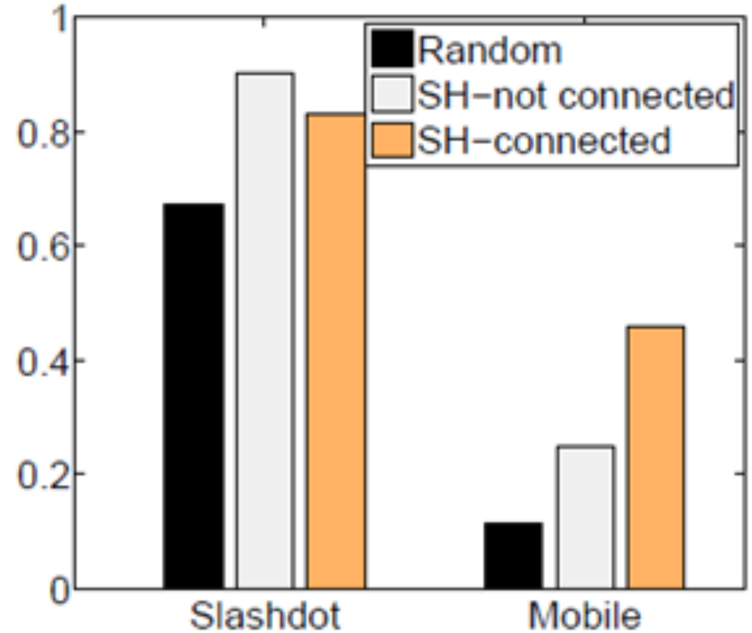
- Community kernel detection
  - GOAL : obtain the importance of each **node** within each **community** (as **kernel members**).
  - HOW : kernel members are **more** likely to connect structural hole spanners.
  - Clear improvements on F1-score, average of 5%



Model applications

- Link prediction
  - GOAL : predict the types of social relationships (on Mobile and Slashdot)
  - HOW : users are more likely to have the **same type** of relationship with structural hole spanners.

Probabilities that two users (A and B) have the same type of relationship with user C, conditioned on whether user C spans a structural hole or not.



[1] J. Tang, T. Lou, and J. Kleinberg. Inferring Social Ties across Heterogeneous Networks. In **WSDM'12**. pp. 743-752.

Model applications

- Link prediction
  - GOAL : predict the types of social relationships (on Mobile and Slashdot)
  - HOW : users are more likely to have the **same type** of relationship with structural hole spanners.
  - Significantly improvement of 1% to 6%

Dataset	Algorithm	K	Precision	Recall	F1-score
	PFG	-	0.9111	0.5694	0.7008
	PFG(HIS)	5	0.8958	0.5972	0.7166
	PFG(HIS)	15	0.8491	0.6250	0.7200
Mobile	PFG(HIS)	25	0.8519	**0.6389**	**0.7302**
	PFG(MaxD)	5	**0.9130**	0.5833	0.7118
	PFG(MaxD)	15	0.8776	0.5972	0.7107
	PFG(MaxD)	25	0.8723	0.5972	0.7090
	PFG	-	0.6619	0.7281	0.6934
	PFG(HIS)	100	0.6562	0.7965	0.7196
	PFG(HIS)	150	0.6615	**0.8241**	**0.7339**
Slashdot	PFG(HIS)	200	**0.6788**	0.7886	0.7296
	PFG(MaxD)	100	0.6602	0.7542	0.7041
	PFG(MaxD)	150	0.6667	0.7532	0.7073
	PFG(MaxD)	200	0.6619	0.7775	0.7151

[1] J. Tang, T. Lou, and J. Kleinberg. Inferring Social Ties across Heterogeneous Networks. In **WSDM'12**. pp. 743-752.

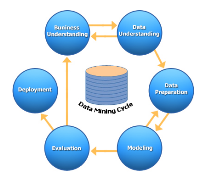
Conclusion

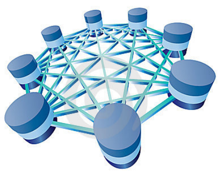
Conclusion

- Study an interesting problem : structural hole spanner detection.
- Propose two models (HIS and MaxD) to detect structural hole spanner in large social networks, and provide theoretical analysis.
- Results
  - **1%** twitter users control **25%** retweeting behaviors between communities.
  - Application to Community kernel detection and Link prediction

Future works

- Combine the topic leveled information with the user network information.
- Dynamics of structural holes





Artificial Intelligence **Data Mining** **Database**

- What's the difference between the patterns of structural hole spanners on other networks?

Thanks you！

Collaborators: Tiancheng Lou (**Google**)

Jon Kleinberg (**Cornell**),

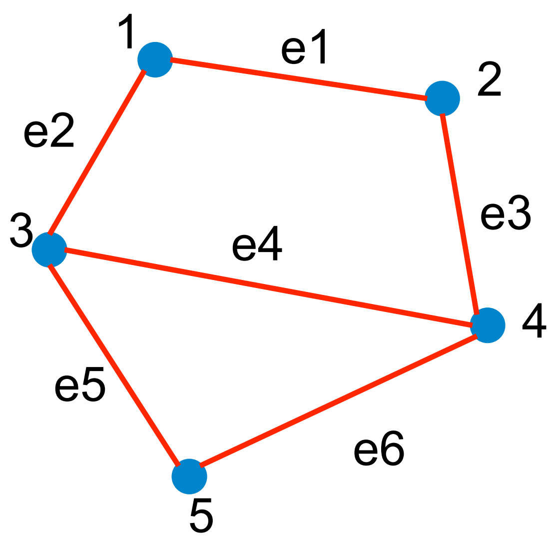
Yang Yang, Cheng Yang (**THU**)

Jie Tang, KEG, Tsinghua U, http://keg.cs.tsinghua.edu.cn/jietang

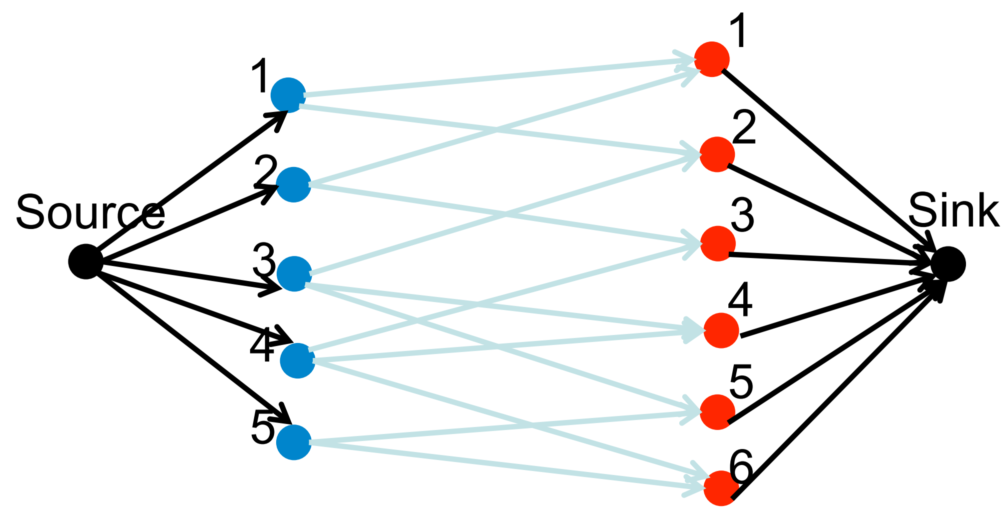
Download data & Codes, http://arnetminer.org/download

Hardness Proof

Instance G = (V, E) of **K-Denest Subgraph**



Minimal node-cut problem

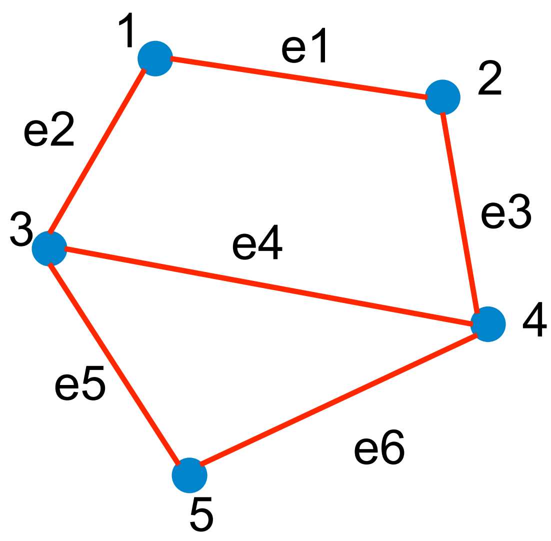


capacity = 1, iff corresponding node exists in the edge (set of 2 nodes)

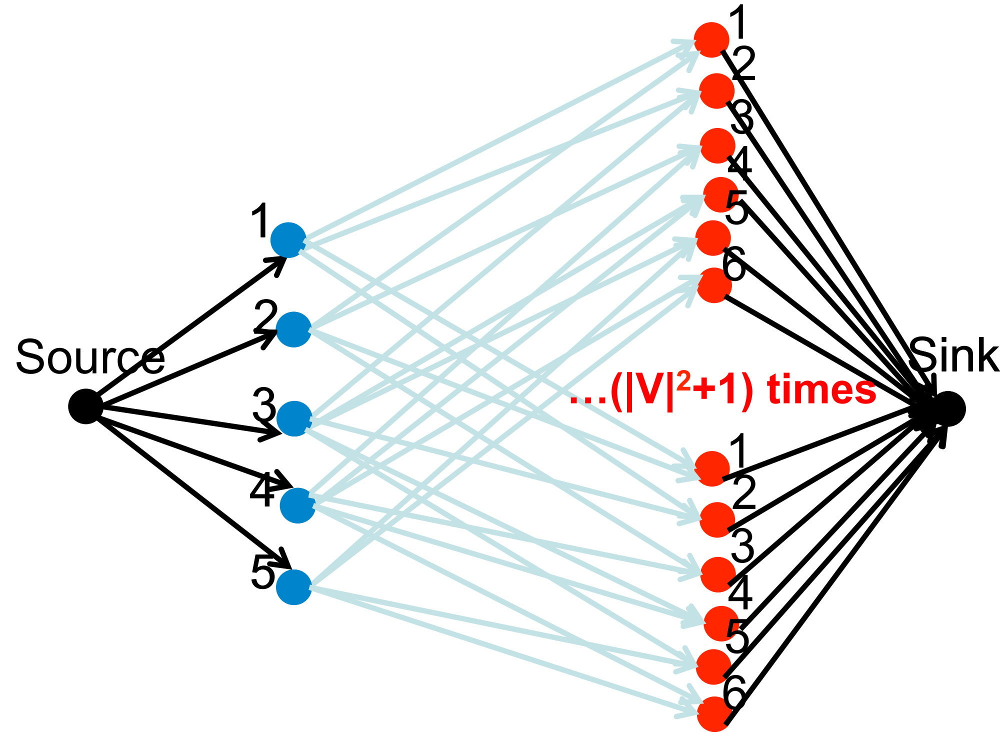
capacity = $(|V|^2 + 1)\ |E|$

Hardness Proof

Instance G = (V, E) of **K-Denest Subgraph**



Minimal node-cut problem



capacity = 1, iff corresponding node exists in the edge (set of 2 nodes)

capacity = $(|V|^2 + 1)\ |E|$

Instance φ is satisfied **iff** there exists a subset $|V_{SH}| = k$, such that $Q(V_{SH}, C) >= d(|V|^2+1)$