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
Core Research in Social Network

Application
- Prediction
- Search
- Information Diffusion
- Advertise

Social Network Analysis
- Macro
  - BA model
  - ER model
- Meso
  - Community
  - Structural behavior
  - Structural hole
- Micro
  - Social tie
  - Action
  - Social influence

Theory
- Social Theories
- Algorithmic Foundations

BIG Social Data
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]

Few People Connect the World

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?

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:** Structural hole spanner vs Opinion leaders

**Challenge 2:** Who controls the information diffusion?

82 overlapped PC members of SIGMOD/ICDT/VLDB and SIGKDD/ICDM during years 2007 – 2009.
Mining Top-k Structural Hole Spanners

Problem Definition

Which node is the best structural hole spanner?

Community 1

Community 2

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

• INPUT:
  – A social network, \( G = (V, E) \) and \( L \) communities \( C = (C_1, C_2, ..., C_L) \)

• 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

<table>
<thead>
<tr>
<th></th>
<th>#User</th>
<th>#Relationship</th>
<th>#Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coauthor</td>
<td>815,946</td>
<td>2,792,833</td>
<td>1,572,277 papers</td>
</tr>
<tr>
<td>Twitter</td>
<td>112,044</td>
<td>468,238</td>
<td>2,409,768 tweets</td>
</tr>
<tr>
<td>Inventor</td>
<td>2,445,351</td>
<td>5,841,940</td>
<td>3,880,211 patents</td>
</tr>
</tbody>
</table>

- 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**”?
The two-step information flow theory\cite{Katz:1973} suggests structural hole spanners are connected with many “opinion leaders”.

Results: Opinion leaders control 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.

- Structural hole spanners control the information diffusion between communities.
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
  \[ I(v, C_i) = \max \{ I(v, C_i), \alpha I(u, C_i) + \beta S H(u, S) \} \]
  \[ H(v, S) = \min \{ I(v, C_i) \} \]

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

\( \alpha \) and \( \beta \) are two parameters.
Algorithm for HIS

Input: $G = (V, E)$, parameters $\alpha_i$, $\beta_S$, and convergence threshold $\epsilon$

Output: Importance $I$ and structural hole score $H$

Initialize $I(v, C_i)$ according to Eq. 4:

```
repeat
  foreach $v \in V$ do
    foreach $C_i \in C$ do
      $P(v, C_i) = \max_{S \subseteq C \land C_i \subseteq S} \{ \alpha_i I(v, C_i) + \beta_S H(v, S) \}$
    end
  end
  foreach $v \in V$ do
    foreach $C_i \in C$ do
      $I'(v, C_i) = \max \{ I(v, C_i), \max_{u \in V, u \neq v} P(u, C_i) \}$
    end
    foreach $S \subseteq C$ do
      $H'(v, S) = \min_{C_i \in S} I'(v, C_i)$
    end
  end
  Check the $\epsilon$-convergence condition by
  $$\max_{v \in V, C_i \in C} |I'(v, C_i) - I(v, C_i)| \leq \epsilon$$
  Update $I = I'$ and $H = H'$;
until Convergence;
```

$I(v, C_i) = r(v), \quad v \in C_i$

$I(v, C_i) = 0, \quad v \notin C_i$

By PageRank or HITS

Parameter to control the convergence
Theoretical Analysis—Existence

- Given \( \alpha_i \) and \( \beta_S \), 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 only if direction
  - Suppose \( \alpha_i + \beta_S > 1 \), \( S = \{C_{\text{blue}}, C_{\text{yellow}}\} \)
  - \( r(u) = r(v) = 1 \);
  - \( I(u, C_{\text{blue}}) = I(u, C_{\text{yellow}}) = 1 \);
  - \( H(u, S) = \min \{ I(u, C_{\text{blue}}), I(u, C_{\text{yellow}}) \} = 1 \);
  - \( I(v, C_{\text{yellow}}) \geq \alpha_i I(u, C_i) + \beta_S H(u, S) = \alpha_i + \beta_S > 1 \)

\[ I(v, C_i) = \max \{ I(v, C_i), \alpha_i I(u, C_i) + \beta_S H(u, S) \} \]
\[ H(v, S) = \min \{ I(v, C_i) \} \]
Theoretical Analysis—Existence

• Given $\alpha_i$ and $\beta_S$, 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
    \[
    I^{(k+2)}(v, C_i) = \alpha_i I^{(k+1)}(u, C_i) + \beta_S H^{(k+1)}(u, S)
    \leq \alpha_i [I^{(k)}(u, C_i) + \gamma^k] + \beta_S [H^{(k+1)}(u, S) + \gamma^k]
    \leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k+1)}(u, S) + \gamma^{k+1}
    \leq I^{(k+1)}(u, C_i) + \gamma^{k+1}
    \]
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

Removing $V_6$ decreases the minimal cut as 2
Model Two: MaxD

- Structural holes spanners play an important role in information diffusion

\[ Q(V_{SH}, C) = \text{MC}(G, C) - \text{MC}(G \setminus V_{SH}, C) \]

\( \text{MC}(G, C) = \) the minimal cut of communities C in G.
Hardness Analysis

\[ Q(V_{SH}, C) = MC (G, C) - MC (G \setminus V_{SH}, C) \]

• 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 $MC(G\setminus V_{SH}) \leq n(m-d)$
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 \( MC(G \setminus V_{SH}) \leq n \cdot (m-d) \).
Proof: NP-hardness (cont.)

• For the *if* direction
  – If there exists a k-subset $V_{SH}$ such that $\text{MC}(G\backslash V_{SH}) \leq n*(m-d)$;
  – Denote $V_{SH'} = V_{SH} \cup \{x\}$, the size of $V_{SH'}$ is at most k, and $\text{MC}(G\backslash V_{SH'}) \leq n*(m-d)$;
  – 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 $V_8$ and decrease the minimal cut from 7 to 4
Step 2: select $V_6$ and decrease the minimal cut from 4 to 2
Step 3: select $V_{12}$ 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**: $G = (V, E), k, l, C = \{C_i\}$

**Output**: Top-$k$ structural hole nodes $V_{SH}$

Initialize $V_{SH} = \emptyset$

while $|V_{SH}| < k$

  Initialize $f(v) = 0$, for each $v \in V$

  foreach non empty $S \subset \{1, \cdots, l\}$ do

  \[ E_S = \cup_{i \in S} C_i \text{ and } E_T = \cup_{i \not\in S} C_i \]

  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

  \[ \text{Add } f(v) \text{ by the flow through node } v \]

  end

  Choose $O(k)$ nodes with the largest $f$ as candidates $D$

  Compute $p^* = \arg \min_{p \in D} MC(G \setminus (V_{SH} \cup \{p\}), C)$

  Update $V_{SH} = V_{SH} \cup \{p^*\}$

end

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

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

**Complexity**: $O(2^lT_2(n))$; $T_2(n)$—the complexity for computing min-cut

**Approximation ratio**: $O(\log l)$
Results
Experiment

<table>
<thead>
<tr>
<th></th>
<th>#User</th>
<th>#Relationship</th>
<th>#Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coauthor</td>
<td>815,946</td>
<td>2,792,833</td>
<td>1,572,277 papers</td>
</tr>
<tr>
<td>Twitter</td>
<td>112,044</td>
<td>468,238</td>
<td>2,409,768 tweets</td>
</tr>
<tr>
<td>Inventor</td>
<td>2,445,351</td>
<td>5,841,940</td>
<td>3,880,211 patents</td>
</tr>
</tbody>
</table>

- 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
  – +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

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 job.

- 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.

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

- Running time of different algorithms in three data sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Pathcount</th>
<th>2-Step</th>
<th>PageRank</th>
<th>HIS</th>
<th>MaxD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coauthor</td>
<td>350.66s</td>
<td>4.71s</td>
<td>0.20s</td>
<td>0.60s</td>
<td>189.78m</td>
</tr>
<tr>
<td>Twitter</td>
<td>32.03m</td>
<td>12.09s</td>
<td>0.67s</td>
<td>3.87s</td>
<td>602.37m</td>
</tr>
<tr>
<td>Inventor</td>
<td>494.3 hr</td>
<td>98.96s</td>
<td>3.61s</td>
<td>26.11s</td>
<td>370.8hr</td>
</tr>
</tbody>
</table>

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.

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.

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%

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>$K$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PFG</td>
<td>-</td>
<td>0.9111</td>
<td>0.5694</td>
<td>0.7008</td>
</tr>
<tr>
<td></td>
<td>PFG(HIS)</td>
<td>5</td>
<td>0.8958</td>
<td>0.5972</td>
<td>0.7166</td>
</tr>
<tr>
<td></td>
<td>PFG(HIS)</td>
<td>15</td>
<td>0.8491</td>
<td>0.6250</td>
<td>0.7200</td>
</tr>
<tr>
<td>Mobile</td>
<td>PFG(HIS)</td>
<td>25</td>
<td>0.8519</td>
<td>0.6389</td>
<td>0.7302</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>5</td>
<td>0.9130</td>
<td>0.5833</td>
<td>0.7118</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>15</td>
<td>0.8776</td>
<td>0.5972</td>
<td>0.7107</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>25</td>
<td>0.8723</td>
<td>0.5972</td>
<td>0.7090</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td></td>
<td>0.6619</td>
<td>0.7281</td>
<td>0.6934</td>
</tr>
<tr>
<td></td>
<td>PFG(HIS)</td>
<td>100</td>
<td>0.6562</td>
<td>0.7965</td>
<td>0.7196</td>
</tr>
<tr>
<td>Slashdot</td>
<td>PFG(HIS)</td>
<td>150</td>
<td>0.6615</td>
<td>0.8241</td>
<td>0.7339</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>100</td>
<td>0.6602</td>
<td>0.7542</td>
<td>0.7041</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>150</td>
<td>0.6687</td>
<td>0.7532</td>
<td>0.7073</td>
</tr>
<tr>
<td></td>
<td>PFG(MaxD)</td>
<td>200</td>
<td>0.6619</td>
<td>0.7775</td>
<td>0.7151</td>
</tr>
</tbody>
</table>

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

• 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,
Download data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://arnetminer.org/download
Hardness Proof

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

capacity = 1, iff corresponding node exists in the edge (set of 2 nodes)
capacity = $(|V|^2 + 1) |E|$
Instance $G = (V, E)$ of **K-Denest Subgraph**

**Hardness Proof**

**Minimal node-cut problem**

- Capacity $= 1$, iff corresponding node exists in the edge (set of 2 nodes)
- Capacity $= (|V|^2 + 1) |E|

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