Scalable Mining of Massive Networks: Distance-based Centrality, Similarity, and Influence

Edith Cohen
Tel Aviv University
Graph Datasets:
Represent relations between “things”

Bowtie structure of the Web Broder et. al. 2001

Dolphin interactions
Graph Datasets

- Hyperlinks (the Web)
- Social graphs (Facebook, Twitter, LinkedIn, ...): friend, follow, like
- Email logs, phone call logs, messages
- Commerce transactions (Amazon, eBay)
- Road networks
- Communication networks
- Protein interactions
- ...
Analytics on Graphs

Centralities/Influence

- The power/importance/coverage of a node or a set of nodes
- Applications: ranking, viral marketing,...

Similarities/Communities

- How tightly related are 2 or more nodes
- Applications: Friend/Product Recommendations, attribute completion, Advertising, Prediction
Challenges

- **Modeling:**
  - Models must capture “reality”
  - Be flexible to allow fitting to problem (tunable parameters)

- **Scalability:**
  - $10^{9+}$ interactions (edges), $10^{8+}$ entities (nodes)
Centrality, Similarity, Influence
Structural measures (based on the set of interactions)

**Local measures**: only depend on the set of neighbors (paths of length 1)
- Degree centrality,
- Influence as cardinality of union of neighbors
- Similarity by relation of neighbor sets (Jaccard, Adamic/Adar)

**Advantages:**
- Gets the “first order bit”
- Scalable

**Disadvantages:**
- Limited recall
- Spammable
Centrality, Similarity, Influence
Structural measures (based on the set of interactions)

**Global Measures**: depend on the paths ensemble

- **Higher recall** but **much less scalable**

- Random walk based (Geometric/Heat Kernel): Centrality, similarity, influence thru (personalized) page rank

- **Katz** (Sum of paths, discounted by length)

- **Resistance**: Similarity by effective resistance

- **Distance/reachability based Measures**: centrality, similarity, influence
Distance/Reachability based measures of Centrality, Influence, Similarity

- **Closeness centrality:** Ability of a node to “reach” other nodes [Bavelas 1950]+++:

- **Influence:** Ability of a set of nodes to reach others [GLM2001,KKT2003,] [Gomez-Rodriguez+, ICML 2011], [Du+ NIPS 2013]. [C’DPW 2014] +++:

- **Closeness similarity:** Relates two nodes based on the similarity of their “reach” [C’DFGGGW2013], (local: [AA 2005,LK2007]):
This talk: Unified treatment of distance/reachability-based measures

- **Models:**
  - Basic definition of “distance-based”
  - Different ways of using distance/reachability
  - Argue that models capture intuitive properties
- **Scalability:** Sketching and estimation
Distance vectors

SP distances:

0 5 6 7 10 10 13 14 15 15 16 17 17

Diagram of a network with distances labeled on the edges.
Distance vectors

SP distances:

17 17 16 13 7 17 16 13 16 3 1 0 4
Distance vectors

SP distances $d_{ij}$:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>10</th>
<th>10</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>7</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

```plaintext
0 0 0 0 0 0 0 0
```

```plaintext
0 0 0 0 0 0 0 0
```

```plaintext
0 0 0 0 0 0 0 0
```
Distance-based measures

SP distances $d_{ij}$:

\[
\begin{array}{cccccccccccccc}
0 & 5 & 6 & 7 & 10 & 10 & 13 & 14 & 15 & 15 & 16 & 17 & 17 \\
17 & 17 & 16 & 13 & 7 & 17 & 16 & 13 & 16 & 3 & 1 & 0 & 4 \\
\end{array}
\]

Relate entities based on their vectors.

- Closeness centrality $C(\text{ })$
- Closeness Similarity $\text{Sim}(\text{ }, \text{ }, \text{ })$
- Influence $\text{Inf}(\text{ }, \text{ }, \text{ }, \text{ })$
Distance-based Models

SP distances $d_{ij}$:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>10</th>
<th>10</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>7</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity  
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances
Distance-based Models

SP distances $d_{ij}$:

0 5 6 7 10 10 13 14 15 15 16 17 17
17 17 16 13 7 17 16 13 16 3 1 0 4

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances
Topic Filter $\beta(i)$

Weigh entities (nodes or entries of vectors) according to:
- Topic, interests, education level, age, community, geography, language, product type,…
- Applications: focus measure on topic. Useful for recommendations, attribute completion, targeted ads

Similar to the role of “start state” in PPR

“Evil”

$\beta$: 1 0 0 0.8 0.9 0 0.1
Distance-based Models

SP distances $d_{ij}$:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>10</th>
<th>10</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>7</td>
<td>17</td>
<td>16</td>
<td>13</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

- Inbound/outbound/undirected distance
- Distance/Reachability/Dijkstra (NN) rank
- Topic filter $\beta(v) \geq 0$: weighted relevance of entity
- Distances vs. decayed distances $\alpha(d) \geq 0$
- Robustness by randomizing lengths/presence of edges (REL), or using multiple “example” instances
Farness-penalty vs. Closeness-reward

- To emphasize **penalty** for weak coverage of **far nodes**, we use distances as is: The norm of the distances vector of a node is dominated by “far” nodes. *As in classic closeness centrality, facility location, k-medians/means*

- To emphasize **reward** for better coverage of **close nodes**, we apply a decay (“kernel”) function $\alpha(d)$ to the distance: The norm is dominated by closer nodes. *As in influence models and local measures.*

- Exact computation: both requires full distance vectors.
- Scalable approximation: different techniques
Quantifying Closeness Reward: Distance Decay (Kernel) Functions

Specify how importance/relevance decays with distance

\[ \alpha(d) = \begin{cases} 
\exp(-\lambda d) & \text{(exponential)}, \\
\frac{1}{d} & \text{(geometric)}, \\
\exp(-\lambda d^2) & \text{(Gaussian)} \\
d < \infty? \ 1: 0 & \text{(reachability)} \\
d < T? \ 1: 0 & \text{(threshold)} 
\end{cases} \]

Similar to the role of kernel distribution in PPR
### Closeness Vectors

**SP distance vectors:**

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>22</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>16</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>15</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>10</td>
</tr>
</tbody>
</table>

**Closeness vectors:**

\[
\alpha(d_{ij}) = \frac{1}{1 + d_{ij}} \quad \beta \equiv 1
\]

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{1}{21})</td>
<td>(\frac{1}{13})</td>
<td>(\frac{1}{23})</td>
<td>(\frac{1}{18})</td>
<td>(\frac{1}{18})</td>
<td>(\frac{1}{16})</td>
<td>(\frac{1}{17})</td>
<td>(\frac{1}{13})</td>
<td>(\frac{1}{3})</td>
<td>1</td>
<td>(\frac{1}{4})</td>
<td>(\frac{1}{5})</td>
<td>(\frac{1}{8})</td>
</tr>
<tr>
<td>(\frac{1}{14})</td>
<td>(\frac{1}{15})</td>
<td>(\frac{1}{16})</td>
<td>(\frac{1}{11})</td>
<td>1</td>
<td>(\frac{1}{7})</td>
<td>(\frac{1}{6})</td>
<td>(\frac{1}{8})</td>
<td>(\frac{1}{16})</td>
<td>(\frac{1}{18})</td>
<td>(\frac{1}{17})</td>
<td>(\frac{1}{18})</td>
<td>(\frac{1}{11})</td>
</tr>
</tbody>
</table>
Closeness Centrality

- Classic (farness penalty)
  \[ C(i) = (n - 1)/\sum_{j} d_{ij} \beta(j) \]

- Distance-decay (closeness reward)
  \[ C(i) = \sum_{j} \alpha(d_{ij}) \beta(j) \]
Closeness Matrix
(distance decay+topic filter)

\[ c_{ij} = \alpha(d_{ij}) \beta(j) \]

\( c_i \) (row \( i \)) is the **closeness vector** of node \( i \)

**Centrality:**

\[ C(i) = \sum_{j} c_{ij} \]

**Influence:**

of a set of nodes

\[ \text{Inf}(S) = \sum_{j} \max_{i \in S} c_{ij} \]
### Example: (dist-decay) Closeness Centrality

#### SP distance vectors:

- (Dist-1) distance decay
- Closeness vectors:

\[
\alpha(d_{ij}) = \frac{1}{1 + d_{ij}} \quad \beta = 1
\]

#### Closeness vectors:

- SP distance vectors:
  - 20 12 22 17 17 15 16 12 2 0 3 4 7
  - 13 14 15 10 0 6 5 7 15 17 16 17 10

- Closeness vectors:
  - 1/21 1/13 1/23 1/18 1/18 1/16 1/17 1/13 1/3 1 1/4 1/5 1/8
  - 1/14 1/15 1/16 1/11 1 1/7 1/6 1/8 1/16 1/18 1/17 1/18 1/11

- \(C(\text{dark green}) \approx 2.05\)  
- \(C(\text{green}) \approx 2.11\)
Example + topic filter: Closeness to Evil

\[ \alpha(d_{ij}) = \frac{1}{1+d_{ij}} \quad \beta \equiv \text{evilness} \]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
0.8 & 0.1 & 0.1 & 0.8 & 1 & 0.1 & 0 & 0 & 0 & 0.1 & 0 & 0.9 \\
\end{array}
\]
Example: Influence from Vectors

### SP distance vectors:

\[
\begin{array}{cccccccccccccc}
20 & 12 & 22 & 17 & 17 & 15 & 16 & 12 & 2 & 0 & 3 & 4 & 7 \\
13 & 14 & 15 & 10 & 0 & 6 & 5 & 7 & 15 & 17 & 16 & 17 & 10 \\
\end{array}
\]

### Closeness vectors:

\[
\begin{array}{cccccccccccccc}
\frac{1}{21} & \frac{1}{13} & \frac{1}{23} & \frac{1}{18} & \frac{1}{18} & \frac{1}{16} & \frac{1}{17} & \frac{1}{13} & \frac{1}{3} & 1 & \frac{1}{4} & \frac{1}{5} & \frac{1}{8} \\
\frac{1}{14} & \frac{1}{15} & \frac{1}{16} & \frac{1}{11} & 1 & \frac{1}{7} & \frac{1}{6} & \frac{1}{8} & \frac{1}{16} & \frac{1}{18} & \frac{1}{17} & \frac{1}{18} & \frac{1}{11} \\
\end{array}
\]

\[
\alpha(d_{ij}) = \frac{1}{1+d_{ij}} \quad \beta \equiv 1
\]

\[
\text{Inf}(\text{Vector}) \approx 3.64
\]

Sum of max (Submodular)

\[
C(\text{Vector}) \approx 2.11
\]

\[
C(\text{Vector}) \approx 2.05
\]
Closeness Similarity: Local Measures

**Local:** $\text{sim}(i, j)$ depends only on immediate neighbors of $i, j$

**Jaccard:**
\[
J(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}
\]

**Adamic Adar:**
\[
\text{AA}(i, j) = \frac{1}{\log |\Gamma(i) \cap \Gamma(j)|}
\]

<table>
<thead>
<tr>
<th>Graph</th>
<th>$J(i, j)$</th>
<th>$\text{AA}(i, j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Graph 1" /></td>
<td>$\frac{2}{9}$</td>
<td>$\frac{1}{\log 5}$</td>
</tr>
<tr>
<td><img src="image2" alt="Graph 2" /></td>
<td>$\frac{3}{5}$</td>
<td>$\frac{3}{\log 4}$</td>
</tr>
</tbody>
</table>

Work well for $d_{ij} \leq 2$, but no recall for $d_{ij} > 2$
Closeness Similarity: Global Measures
Based on full closeness vectors

\[ c_{ij} = \alpha(d_{ij}) \beta(j) \]

**weighted Jaccard:**

\[ \text{sim}(i, j) = \frac{\sum_h \beta(h) \min c_{ih}, c_{jh}}{\sum_h \beta(h) \max c_{ih}, c_{jh}} \]

**Cosine:**

\[ \text{sim}(i, j) = c_i \cdot c_j. \]

**\(L_p\) Distance:**

\[ L_p\text{-dist}(i, j) = \|c_i - c_j\|_p \]
Example: (global) Closeness Similarity

SP distance vectors:

| 20 | 12 | 22 | 17 | 17 | 15 | 16 | 12 | 2 | 0 | 3 | 4 | 7 |
| 13 | 14 | 15 | 10 | 0 | 6 | 5 | 7 | 15 | 17 | 16 | 17 | 10 |

Closeness vectors:

| 1/21 | 1/13 | 1/23 | 1/18 | 1/18 | 1/16 | 1/17 | 1/13 | 1/3 | 1 | 1/4 | 1/5 | 1/8 |
| 1/14 | 1/15 | 1/16 | 1/11 | 1 | 1/5 | 1/6 | 1/8 | 1/16 | 1/18 | 1/17 | 1/18 | 1/11 |

Weighted Jaccard: \( \frac{\sum_x \min(C_{ix}, C_{jx})}{\sum_x \max(C_{ix}, C_{jx})} \approx 0.21 \)

Cosine similarity: \( \frac{\sum_x C_{ix} C_{jx}}{\|C_i\|_2 \|C_j\|_2} \approx 0.18 \)

\( L_1 \) Distance: \( \sum_x |C_{ix} - C_{jx}| \approx 2.8 \)

\[ \alpha(d_{ij}) = \frac{1}{1+d_{ij}} \]

\[ \beta \equiv 1 \]
So far: Definitions and some motivation of distance-based measures

Next: Scalability thru Sketches and estimators
Scalable Closeness Centrality

- Closeness centrality of a node $v$ can be computed exactly using a SSSP computation from $v$.
- But this takes near-linear time per node: Does not scale when we want centralities of many or all nodes.

We want to estimate all centralities within a small relative error using near-linear computation.

- Classic (farness penalty): [C’DPW : COSN 2014]
- Previously additive error: [EW SODA 2001, OCL 2008]
- Distance-decay (closeness reward): All-distances sketches [C’ 94], [C’ Kaplan SIGMOD 2004] [C’ : PODS 2014].
Scaling Up (closeness reward)

- All-distances and reachability node sketches (ADS) [C’ 94], [C’ Kaplan 2004] [C’ : PODS 2014].

- Distance-decay closeness centrality [C’ 94], [C’ Kaplan 2004] [Boldi Vigna 2013] [C’ : PODS 2014].
- Closeness similarity [CDFGGGW: COSN 2013]
- Influence computation and maximization [CWY KDD 2009] [C’ Delling Pajor Werneck: CIKM 2014] [Du Song Gomez-Rodruiguez Zha NIPS 2013] [C’ Delling Pajor Werneck 2015]
All-Distances Sketches (ADS) [C’ 94] + per-node summary structures of Un/Directed, Un/Weighted networks

For a node $i \in [n]$ : $\text{ADS}(i)$ “samples” the distance vector of $i$

$m$ edges, $n$ nodes, parameter $k \geq 1$ trades-off sketch size and information
All-Distances Sketches: Definition

\( \text{ADS}(\nu) \) is a list of pairs of the form \((i, d_{vi})\)

- Draw a random permutation of the nodes:
  \[ r : [n] \to [n] \]
- \( i \in \text{ADS}(\nu) \iff r(i) < k^{\text{th}} \) smallest rank amongst nodes that are closer to \( \nu \) than \( i \)

This is a \textit{bottom-k ADS}, there are other ADS “flavors”
**All-Distances Sketches: Definition**

\( \text{ADS}(v) \) is a list of pairs of the form \((i, d_{vi})\)

- Draw a random permutation of the nodes:
  \[ r : [n] \rightarrow [n] \]

- \( i \in \text{ADS}(v) \iff r(i) < k^{\text{th}} \text{ smallest rank amongst nodes that are closer to } v \text{ than } i \)

Expected size of \( \text{ADS}(i) \) is \( \sum_{j=1}^{n} \min \left\{ 1, \frac{k}{j} \right\} < k \ln n \)

- The \( j^{\text{th}} \) closest node to \( i \) is included with probability \( \min \{1, \frac{k}{j}\} \)
Random permutation of nodes
ADS example $k = 1$

All nodes sorted by SP distance from

0.63  0.42  0.56  0.84  0.07  0.35  0.49  0.77  0.91  0.21  0.28  0.14  0.70

$k = 1$:  

0.63  0.42  0.07
ADS example $k = 2$

Sorted by SP distances from

$0.63 \ 0.42 \ 0.56 \ 0.84 \ 0.07 \ 0.35 \ 0.49 \ 0.77 \ 0.91 \ 0.21 \ 0.28 \ 0.14 \ 0.70$

$k = 2:$

$0.63 \ 0.42 \ 0.56 \ 0.07 \ 0.35 \ 0.21 \ 0.14$
Sketch Coordination

We use the same permutation to obtain the ADS of all nodes, as a result:

- ADS Sketches of different nodes are coordinated: related in a way that is useful for queries that involve multiple nodes (similarities, influence, distance) [Brewer, Early, Joyce 1972]
- Generalize MinHash sketches by adding a time/distance dimension
Computing ADSs Efficiently

- Pruned Dijkstra’s Algorithm: Build ADS entries by increasing permutation rank.
- Local (Dynamic programming) Algorithm: Build ADS entries by increasing distance.

Either way, we scan outgoing edges of a node only after its ADS is modified ⇒ number of edge traversals is bounded by ADS size: $O(km \ln n)$
Computing ADSs Efficiently

Perform pruned Dijkstra from nodes by increasing permutation rank:
Estimation from sketches
Estimation with All-Distances Sketches

- Closeness centrality of node $i$, from $\text{ADS}(i)$ with NRMSE $O(\frac{1}{\sqrt{k}})$ (+concentration) [C’ 94, C’ Kaplan 04, C’ 14]

- Influence of $S$, from $\text{ADS}(i)$ $i \in S$, with NRMSE $O(\frac{1}{\sqrt{k}})$ [C’ 94, Chen WY KDD 09, Du SGZ NIPS 13, C’ DPW 2014 ]

- Closeness similarity $\text{sim}(i, j)$ (W/Jaccard, cosine), from $\text{ADS}(i)$ and $\text{ADS}(j)$ w. RMSE $O(\frac{1}{\sqrt{k}})$ [C’ DFGGW COSN 13]

- $L_p$ distance, from $\text{ADS}(i)$ and $\text{ADS}(j)$ with RMSE $O(\frac{1}{\sqrt{k}})$ times max norm using [C’ KDD 2014]

Side note: ADSs can also be used as distance oracles (estimate pairwise distances), spanners, and more
Historic Inverse Probability (HIP) probability & estimator [C’ 2014]

- The **HIP inclusion probability** $p_{vi}$ of $i \in \text{ADS}(v)$ is *conditioned* on fixed rank values of other nodes.

- To estimate centrality of $v$, we estimate the "presence" of $i$ with respect to $v$: $I_{v \sim i}$ ($=1$ if $v \sim i$, $0$ otherwise) using "inverse-probability" [HT52] $a_{vi} = \frac{1}{p}$. This is unbiased (when $p > 0$).

- For bottom-$k$ ADS and $r \sim U[0,1]$

  
  $$p_{vi} = k^{th}\{r(u) | u \in \text{ADS}(v) \land d_{vu} < d_{vi}\}$$
Example: HIP estimates

Bottom-2 ADS of

\[
\begin{array}{cccccccc}
0.63 & 0.42 & 0.56 & 0.07 & 0.35 & 0.21 & 0.14 \\
\end{array}
\]

\[
p: \quad 1 & 1 & 0.63 & 0.56 & 0.42 & 0.35 & 0.21 \\
\]

\[
a = \frac{1}{p}: \quad 1 & 1 & 1.59 & 1.79 & 2.38 & 2.86 & 4.76 \\
\]

\[
p:2^{nd} \text{ smallest } r \text{ value among closer nodes}
\]
**HIP cardinality estimate**

Bottom-2 ADS of

distance: 0 5 6 10 10 15 17

$a$: 1 1 1.59 1.79 2.38 2.86 4.76

Query: $n_6(v) = \left| \{ i \mid d_{vi} \leq 6 \} \right| = \sum_i I_{d_{vi}\leq6}$

$n_6(v) = \sum_{(i,d_{vi})\in ADS(v) \mid d_{vi}\leq6} a_{vi} = 1 + 1 + 1.59 = 3.59$
Quality of HIP cardinality Estimate

Lemma: The HIP neighborhood cardinality estimator

\[ n_d(v) = \sum_{(i,d_{vi}) \in ADS(v) \mid d_{vi} \leq d} a_{vi} \]

Has Coefficient of Variation

\[ \frac{\sigma}{\mu} \leq \frac{1}{\sqrt{2k-2}} \]

This is \( \sqrt{2} \) improvement over “MinHash” estimators [C’ 94]. Also optimally uses sketch.
HIP estimates of Centrality

\[ C_v = \sum_{i} \alpha(d_{vi}) \beta(i) \]

- \( \alpha \) non increasing; \( \beta \) some filter

\[ \hat{C}_v = \sum_{i \in ADS(v)} a_{vi} \alpha(d_{vi}) \beta(i) \]
HIP estimates: closeness to good/evil

Bottom-2 ADS of

<table>
<thead>
<tr>
<th>distance</th>
<th>0</th>
<th>5</th>
<th>6</th>
<th>10</th>
<th>10</th>
<th>15</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$:</td>
<td>1</td>
<td>1</td>
<td>1.59</td>
<td>1.79</td>
<td>2.38</td>
<td>2.86</td>
<td>4.76</td>
</tr>
<tr>
<td>$\beta$:</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.1</td>
<td>1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Filter: $\beta \in [0,1]$ measures “goodness”

Distance-decay “kernel”: $e^{-d_{vi}}$

$$\widehat{C}_v = \sum_{i \in ADS(v)} a_{vi}\beta(i)e^{-d_{vi}} = e^{-5} + 1.59e^{-6} + 0.2e^{-10} + \cdots$$
Similarity/Influence Estimation

We work with HIP inclusions and sum estimators (estimate separately contribution of each node)

We use *monotone estimation formulation* [C’K Random 2014; C’ PODC 2014] and the *L*\(^*\) *estimator* (unique optimal monotone unbiased nonnegative sum estimator)

- **Influence**: can be estimated by merging ADS sketches and applying centrality estimator, but L\(^*\) is tighter
- **Similarity**: inverse-probability on joint inclusion may not even apply, L\(^*\) gets around the problem
Estimating Jaccard Closeness Similarity

SP-distance; using kernel $\alpha(d_{ij}) = \frac{1}{1+d_{ij}}$:

$$\text{sim}(i, j) = \frac{\sum_x \min C_{ix}, C_{jx}}{\sum_x \max C_{ix}, C_{jx}} = \frac{\sum_x \min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}}{\sum_x \max \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}}$$

- Sum estimator: For all $x$ obtain unbiased estimates $\widehat{N_x}$ of $\min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$; $\widehat{X_x}$ of $\max \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$;

$$\widehat{\text{sim}} (i, j) = \frac{\sum_x \widehat{N_x}}{\sum_x \widehat{X_x}}$$

- $\widehat{X_x} > 0$ or $\widehat{N_x} > 0$ only when $x \in ADS(i) \cup ADS(j)$
  $\Rightarrow$ easy to compute $\widehat{\text{sim}} (i, j)$ using only $ADS(i) \cup ADS(j)$
- We use $L^*$ for $X_x$. We show $\widehat{N_x}$ (inverse probability). Get relative error $\frac{1}{\sqrt{k}}$ on denominator, additive error of same order on numerator.
Estimating Jaccard Closeness Similarity

Estimate $\hat{N}_x$ for $\min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$:

For $x \in ADS(i) \cap ADS(j)$
- We compute HIP probability $p$ for inclusion in intersection
- An inverse probability estimate:
  $$\hat{N}_x = \frac{1}{p} \min \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$$

Otherwise, estimate is $\hat{N}_x = 0$

This is the best we can do because if $x$ is not in intersection, data is consistent with 0 value $\Rightarrow$ any nonzero unbiased estimator must return 0
Estimating Jaccard Closeness Similarity

Estimate $\widehat{X}_x$ for $\max \frac{1}{d_{ix}+1}, \frac{1}{d_{jx}+1}$

For $x \in ADS(i) \cup ADS(j)$

- We can never determine $X_x$ if $x$ is only reachable from one of $i, j$. We can only lower bound $X_x$.
- The L* estimator [C’ 2014] is an optimal extension of inverse probability to such a setting.
NN-rank based closeness similarity

$$\text{sim}(i, j) = \frac{\sum_x \min \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}}{\sum_x \max \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}}$$

$$\pi_{ix}$$: Dijkstra (NN) rank of $$x$$ with respect to $$i$$

Lemma:

$$\text{sim}(i, j)$$ is approximated well by

$$\frac{|\text{ADS}(i) \cap \text{ADS}(j)|}{|\text{ADS}(i) \cup \text{ADS}(j)|}$$

“proof:”

- $$\Pr[x \in \text{ADS}(i) \cap \text{ADS}(j)] \approx k \min \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}$$
- $$\Pr[x \in \text{ADS}(i) \cup \text{ADS}(j)] \approx k \max \frac{1}{\pi_{ix}}, \frac{1}{\pi_{jx}}$$

ADSs can be stored very compactly: no distances, hash of nodes
Next:
Enhancing distance/reachability-based models with REL

REL: randomizing lengths/presence of edges

- (Implicit) The Independent Cascade Diffusion model [Kempe Klienberg Tardos KDD 2003]
- Closeness similarity [C’DFGGW 2013]
Strength of relation between two entities

Basic Intuitions for dependence on path ensemble:

- Increase with shorter paths
- Increase with more (redundant paths)
Boosting distance by Randomizing Edge Lengths (REL)

We expect the strength of a relation to
- \textbf{Increase} with the multiplicity of paths
- \textbf{Decrease} with the length of paths

- “Eigenvector” measures:
  - Rooted PageRank, RWR, Hitting Time, Commute Time, Eff.
  - Resistance, Katz

- SP distances: \textit{weakly}

\

\textbf{strictly}
Randomizing Edge Lengths (REL)

- Replace edge lengths (or presence) with independent random variables
- Look at the *expected* measure
Randomizing Edge Lengths (REL)

**Expected distance** is lower with multiple paths: expectation of the minimum < minimum of expectations
Randomizing Edge Lengths (REL)

**Scalability:**

- use Monte Carlo simulations of model to generate fixed instances (graphs)
- Build sketches for multiple instances
Benefits of REL in network analysis

- Measure does reward for paths multiplicity
- Robustness: Measure is not sensitive to small changes in edge weights (edge presence for reachability)
Next:
Some (preliminary) experimental results
Scalability: Timed Influence with REL

Combined ADS sketches for all nodes, obtained from $\ell = 64$ instances generated by assigning independent Exp distributed edge lengths. Sketch parameter $k = 64$

<table>
<thead>
<tr>
<th>Centrality (1 seed)</th>
<th>and Influence (50 seeds) queries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Queries $\alpha(x) = \exp(-10x)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>#nodes</th>
<th>#edges</th>
<th>[h:m]</th>
<th>time $\mu$s</th>
<th>%err</th>
<th>time $\mu$s</th>
<th>%err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slashdot</td>
<td>77K</td>
<td>828K</td>
<td>1:10</td>
<td>46</td>
<td>5.6%</td>
<td>11880</td>
<td>0.4%</td>
</tr>
<tr>
<td>Gowalla</td>
<td>197K</td>
<td>1.9M</td>
<td>3:55</td>
<td>52</td>
<td>3.8%</td>
<td>17100</td>
<td>0.4%</td>
</tr>
<tr>
<td>Twitter F</td>
<td>456K</td>
<td>15M</td>
<td>19:33</td>
<td>51</td>
<td>3.8%</td>
<td>13800</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

[C’ Delling Pajor Werneck 2014]
Closeness Similarity evaluation

[C’ Delling Fuchs Goldberg Goldszmid Werneck  COSN 2013]

Data Sets

<table>
<thead>
<tr>
<th></th>
<th>Nodes x $10^6$</th>
<th>Edges x $10^6$</th>
<th>ADS label size</th>
</tr>
</thead>
<tbody>
<tr>
<td>arXiv</td>
<td>0.4</td>
<td>28.7</td>
<td>37.9</td>
</tr>
<tr>
<td>DBLP</td>
<td>1.1</td>
<td>9.2</td>
<td>39.1</td>
</tr>
<tr>
<td>twitter</td>
<td>29.6</td>
<td>603.9</td>
<td>101.6</td>
</tr>
<tr>
<td>smallworld</td>
<td>1.0</td>
<td>6.0</td>
<td>40.7</td>
</tr>
</tbody>
</table>
Similarity Evaluation

**Spearman coefficient**: correlation between meta-data ranking and similarity measure ranking (1 = perfect match, 0 = random rankings). On pairs selected “uniformly” by “ground-truth” similarity.

<table>
<thead>
<tr>
<th></th>
<th>arXiv</th>
<th>DBLP</th>
<th>Twitter</th>
<th>SmallWord</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adamic-Adar</td>
<td>0.626</td>
<td>0.746</td>
<td>0.548</td>
<td>0.000</td>
</tr>
<tr>
<td>hops</td>
<td>0.752</td>
<td>0.748</td>
<td>0.169</td>
<td>0.767</td>
</tr>
<tr>
<td>SP distance</td>
<td>0.590</td>
<td>0.634</td>
<td>0.140</td>
<td>0.767</td>
</tr>
<tr>
<td>RWR 0.75</td>
<td></td>
<td>0.734</td>
<td></td>
<td>0.286</td>
</tr>
<tr>
<td>RWR 0.50</td>
<td></td>
<td>0.737</td>
<td></td>
<td>0.617</td>
</tr>
<tr>
<td>RWR 0.25</td>
<td></td>
<td>0.740</td>
<td></td>
<td>0.791</td>
</tr>
<tr>
<td>RWR 0.00</td>
<td></td>
<td>0.500</td>
<td></td>
<td>0.915</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.641</td>
<td>0.742</td>
<td>0.613</td>
<td>0.609</td>
</tr>
<tr>
<td>Closeness REL</td>
<td>0.634</td>
<td>0.752</td>
<td>0.649</td>
<td>0.808</td>
</tr>
</tbody>
</table>
Similarity Evaluation

- No clear winner (networks too different)
- Local measures are limited
- RWR has very good recall (with tuning) but computationally expensive
- Closeness (+REL) : robust, good recall, fast queries (microseconds)

<table>
<thead>
<tr>
<th></th>
<th>arXiv</th>
<th>DBLP</th>
<th>Twitter</th>
<th>SmallWord</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adamic-Adar</td>
<td>0.626</td>
<td>0.746</td>
<td>0.548</td>
<td>0.000</td>
</tr>
<tr>
<td>hops</td>
<td>0.752</td>
<td>0.748</td>
<td>0.169</td>
<td>0.767</td>
</tr>
<tr>
<td>SP distance</td>
<td>0.590</td>
<td>0.634</td>
<td>0.140</td>
<td>0.767</td>
</tr>
<tr>
<td>RWR 0.75</td>
<td></td>
<td>0.734</td>
<td>0.286</td>
<td></td>
</tr>
<tr>
<td>RWR 0.50</td>
<td></td>
<td>0.737</td>
<td>0.617</td>
<td></td>
</tr>
<tr>
<td>RWR 0.25</td>
<td></td>
<td>0.740</td>
<td>0.791</td>
<td></td>
</tr>
<tr>
<td>RWR 0.00</td>
<td></td>
<td>0.500</td>
<td></td>
<td>0.915</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.641</td>
<td>0.742</td>
<td>0.613</td>
<td>0.609</td>
</tr>
<tr>
<td>Closeness REL</td>
<td>0.634</td>
<td>0.752</td>
<td>0.649</td>
<td>0.808</td>
</tr>
</tbody>
</table>
Conclusion

Distance/reachability based measures of centrality, influence, and similarity: global measures that are flexible and highly scalable

- Presented a unified treatment
- Scalability through:
  - All-Distances and reachability sketches
  - Estimators applicable to sketches
- Future: Model fitting framework, Evaluations, Algorithms Engineering
Thank you!
Components of my work are joint work with (subsets of): Daniel Delling, Fabian Fuchs, Andrew Goldberg, Moises Goldszmidt, Haim Kaplan, Thomas Pajor, and Renato Werneck