EventRank: A Framework for Ranking Time-Varying Networks

Joshua O’Madadhain and Padhraic Smyth
Computer Science Department
University of California, Irvine
Network-based ranking algorithms

- vertices represent individuals, edges represent events
- rank: “influence”, “authority”, “centrality”, “prestige”
- based on
  - degree
  - betweenness
  - eigenvector
    - PageRank
    - HITS
    - voltage-based
- all assume implicitly that the network is static
A simple event network

- any network ranking algorithm operating on this representation of the events will give the same rank to D and E, and to A and C
A sequence of events
Event-driven ranks and time

- research citation networks
  - researchers gain prestige/influence over time as they publish papers that are heavily cited
  - prestige may decrease if a researcher stops publishing

- email networks
  - participation can change due to deadlines, vacations, changes of status, ...
Existing methods and temporal data

• apply a static ranking algorithm to successive “snapshots” of the data (restricted to an interval)

• assign weights to edges that reflect elapsed time since the associated events occurred

• ...neither of these consider the sequence of events:
Our contribution

- a framework for constructing network ranking measures with the following properties:
  - ranks change over time according to participation in events (or lack thereof)
  - rank changes respect event sequence

- two types of measures for temporal ranking
  - \textbf{transient}: rank at a specific time $t$
  - \textbf{cumulative}: rank encompassing the interval $[t_0, t]$
    - these ranks are comparable to those of existing algorithms
Model: overview

process events in chronological order; non-participants “send” some of their potential to participants

- conserves potential so potential values are comparable across time
- ensures that an event always causes participants’ potential to increase (unless no non-participants, i.e., spam)
- participants gain potential
- non-participants lose potential
Potential flow
Potential flow: static vs. dynamic

- **static ranking**
  - repeated multiplication of potential vector by constant matrix $M$ used to solve equations for potential steady state

- **dynamic ranking**
  - multiplication by a matrix $M_i$ representing $i$th event
    - updates potential values in response to events (no steady state)
  - transient rank: amount of potential present at time $t$
  - cumulative rank: (function of) potential flow over interval $[t_0, t]$
Requirements for temporal ranking

1. comparability across time
2. participation increases rank in proportion to other participants’ rank
3. participating can always increase rank
4. participants’ ranks don’t decrease
5. non-participants’ ranks don’t increase
6. rank value evolution reflects event sequence
Model: details

\[ R_i(c) = \begin{cases} R_{i-1}(c) & \text{if } c \in P_i \\ R_{i-1}(c) \cdot \left(1 - \frac{\alpha_i}{T_{N_{i-1}}} \right) & \text{if } c \notin P_i \end{cases} \]

potential of \( c \) after \( i \)th event

\[ R_0(c) = \frac{1}{|C|} \]

\[ \bar{R}_i(d) = 1 - R_i(d) \]

participants in \( i \)th event

previous potential

the less potential you had, the more you get

total potential transferred by \( i \)th event

total potential of non-participants

the less potential you had, the more you get
Volatility

- $\alpha_i$ characterizes the potential values’ volatility
  - limited by amount available:
    \[ 0 \leq \alpha_i \leq T_{N_i-1} \]
  - how do we define it?

- baseline model:

  \[ \alpha_i = f \cdot T_{N_i-1} \]
Reply model for $\alpha_i$

- refines baseline model by considering event tempos
- a message is considered a **reply** if the sender has recently received a message from any recipient

$$\alpha_i = f \cdot T_{N_{i-1}} \cdot g(\Delta t_s, G) \cdot h(\Delta t_r, H)$$

<table>
<thead>
<tr>
<th>baseline model</th>
<th>frequent senders</th>
<th>slow replies</th>
</tr>
</thead>
</table>

![Graph](image)
Deriving rank from potential flow

- cumulative rank measures
  - sum of potential gained by sending messages ($S_o$)
    - analogous to outdegree or HITS “hub” score
  - sum of potential gained by receiving messages ($S_r$)
    - analogous to indegree or HITS “authority” score
  - sum of potential values at each step ($S_r$)
Data

- 1 million emails, spanning 21 months, for 628 individuals in an organization
- message ID, sender ID, recipient IDs, timestamp
- no access to content or message headers
- organizational hierarchy position known for 378 individuals
  - derived: distance from top, number of subordinates
How can rank values be validated?

- rank in a social network is subjective: no ground truth

- ranks for smaller networks can be validated in part by comparing their results with those of individual surveys
  - not practical for populations that can’t be interviewed, e.g., large populations (how could Google or Yahoo! do it?)
Experiments

• tested baseline and reply models, $f \in \{0.001, 0.01, 0.1, 0.9\}$
  • reply model: $H = 1$ day, $G = 1$ hour
• measurements:
  • relation between ranks and properties of hierarchy
  • comparison of ranking methods (based on fidelity to hierarchy)
  • sensitivity to parameter settings
• focused on cumulative ranking methods (no point of comparison for transient ranking methods)
• methods tested
  • HITS authority, PageRank, weighted indegree, sum of incoming potential ($S_i$), sum of transient ranks ($S_r$)
Rank and hierarchy properties

Baseline Model, $S$, $f=0.9$

Baseline Model (f = 0.9, $S$)

Corr = .47
Rank and hierarchy depth
Rank and number of subordinates
Comparing ranking algorithms

- individuals should have higher rank than subordinates, lower than superordinates
- compare ranking algorithms based on inversions
  - A and H are inverted if rank(A) > rank(H)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>I_R</th>
<th>I_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>HITS (authority)</td>
<td>1.17</td>
<td>48.88</td>
<td>0.88</td>
</tr>
<tr>
<td>PageRank (α = 0.1)</td>
<td>0.80</td>
<td>41.14</td>
<td>0.92</td>
</tr>
<tr>
<td>indegree</td>
<td>0.54</td>
<td>39.38</td>
<td>0.95</td>
</tr>
<tr>
<td>baseline, Si (f = 0.9)</td>
<td>0.41</td>
<td>16.05</td>
<td>0.96</td>
</tr>
<tr>
<td>baseline, Sr (f = 0.9)</td>
<td>0.63</td>
<td>33.33</td>
<td>0.94</td>
</tr>
<tr>
<td>reply, Si (f = 0.001)</td>
<td>0.50</td>
<td>18.94</td>
<td>0.96</td>
</tr>
<tr>
<td>reply, Sr (f = 0.001)</td>
<td>0.55</td>
<td>30.15</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Sensitivity analysis

• what is the effect of varying \( f \) on cumulative ranks?
  • minimal: rank orderings differ by no more than \( 2\% \)

• what is the effect of varying \( G, H \)?
  • moderate: rank orderings vary by as much as \( 20\% \)
Conclusions

- algorithms based on EventRank framework
  - provide a principled means for calculating rankings that respect event sequences

- email-traffic-based ranks can be a good predictor of position in the organizational hierarchy
  - our algorithms generate ranks which are a better fit to organizational hierarchy than those of existing algorithms
Future Work

- additional data sets
- incorporate header and content data (use to improve reply model)
- application to other types of event data, including undirected relations (e.g. meetings, collaborations)
- analysis of transient rank values to automatically discover patterns in relative ranks over time
  - increasing and decreasing trends
  - periodic bursts at certain times of year