Network Growth and Link Prediction Through an Empirical Lens

Qingyun Liu, Shiliang Tang, Xinyi Zhang, Xiaohan Zhao
Ben Y. Zhao, Haitao Zheng

UC Santa Barbara
qingyun_liu@cs.ucsb.edu
A Fundamental Network Problem

- Network dynamics provide deep insights to understand underlying structure and process

- Social network dynamics are heavily influenced by social recommendations
  - Personal Streaming
  - Related Questions
  - Friend Recommendation

- Link prediction: basis for social recommendations
  - Predicts formation of edges on a given network
A Well Solved Problem?

• Many believe link prediction problem is well addressed
  – Success of the sites/applications based on social recommendations
  – Sheer volume of prior literature

• Empirical evidences show link prediction is not that successful
  – Negative user feedback
  – Potential privacy concern
A Well Solved Problem?

• Many believe link prediction problem is well addressed
  – Success of the sites/applications based on social recommendations
  – Sheer volume of prior literature

• Empirical evidences show link prediction is not that successful
  – Negative user feedback
  – Potential privacy concern

We need to really understand how effective current solutions are
Prior Studies

• Many focus on algorithms/models

• One factor: constrained by the availability of data
  – Public datasets are often small
  – Lack detailed dynamics

• A need to understand how prior studies work in more *realistic* datasets
Limitation & Key Questions

- Traces of network dynamics become available from Online Social Network (OSN)
- Focus on: network structure
  - No meta data, *e.g.*, user mobility information, user phone contacts
  - Explore how these fundamental and general algorithms work
- Key questions
  - How far have we come in understanding network growth?
  - What lessons can we draw from the successes (and failures) of existing algorithms?
  - Can we improve existing approaches by leveraging dynamic data?
Methodology of Our Study

• Looked at traces of network growth from 3 large OSNs
  – Facebook, YouTube, Renren

• Covered 18 representative link prediction algorithms
  – Metric-based algorithms
  – Classification-based algorithms

• Proposed a new technique based on temporal dynamics
  – A preprocessing step that complements all existing algorithms
Outline

• Motivation

• Methodology

• Results from Empirical Study

• New Technique based on Temporal Dynamics
Link Prediction Problem

- Detect *hidden* links
  - Partial graph reconstruction

- Predict *future* links
  - New links in the near future
Metric- vs. Classification-based

• Metric-based: quantify and rank potential edges based on specific metrics
  – Higher rank: more likely to form

  ![Diagram of graph structure]

  Metric = Common Neighbors

• Classification-based: treat link prediction as a classification problem

  SVM, Naïve Bayesian, Random Forest


- Properties in graph structure
- Probabilistic models, matrix (tensor) techniques, ...

- Features can be
  - Basic metrics, *e.g.*, node degree
  - Output from other algorithms
Wild Range of Covered Algorithms

- Covered 18 representative algorithms

**Metric-based (14)**

- Structural Metrics (10)
  - *e.g.* Resource Allocation\([1]\): node neighbors
  - Katz score\([2]\): path properties

- More Complex Models (4)
  - *e.g.* Rescal\([3]\): tensor-based

**Classification-based (4)**

- SVM, Logistic Regression, Naïve Bayesian, Random Forests

**Features**: basic metrics + output of 14 metric-based algorithms

**Output**: a ranking of potential edges

---

\[1\] ZHOU, LÜ and ZHANG. Predicting missing links via local information. European Physical Journal B 71, 4(2009), 623–630

\[2\] KATZ. A new status index derived from sociometric analysis. Psychometrika 18, 1 (1953), 39–43

\[3\] NICKEL, TRESP, AND KRIEGEL. A three-way model for collective learning on multi-relational data. ICML 2011
Examples of Algorithms

- **Metric-based: Katz score**
  - Weighted sum over *all* paths between a node pair
  - Short paths weighted more

- **Classification-based: SVM**

![Graph image]

| Path^2 | = 2

# of hops

11/14/16 Qingyun Liu
Examples of Algorithms

• Metric-based: Katz score
  – Weighted sum over all paths between a node pair
  – Short paths weighted more

• Classification-based: SVM

$|\text{Path}_2| = 2 \quad |\text{Path}_3| = 1$
Examples of Algorithms

- Metric-based: Katz score
  - Weighted sum over all paths between a node pair
  - Short paths weighted more

\[ \text{Katz}(B,D) = \beta \cdot 2 + \beta^2 \cdot 1 \]

- Classification-based: SVM
Examples of Algorithms

• Metric-based: Katz score
  – Weighted sum over all paths between a node pair
  – Short paths weighted more

\[ |\text{Path}^2| = 2 \quad |\text{Path}^3| = 1 \]
\[ \text{Katz}(B,D) = \beta \cdot 2 + \beta^2 \cdot 1 \]

• Classification-based: SVM

From classification to ranking
  ▪ Rank by (signed) distance to the hyperplane
  ▪ Predicted positives furthest from the hyperplane: ranked highest

Input features
  ▪ e.g., max node degree in a pair
  ▪ e.g., Katz score

Qingyun Liu
Datasets

- Traces of network growth from 3 large OSNs
  - With detailed timestamps when edges were created

<table>
<thead>
<tr>
<th>Graph</th>
<th>Trace Length</th>
<th>Trace Start</th>
<th>Trace End</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Nodes</td>
<td>Edges</td>
<td>Nodes</td>
</tr>
<tr>
<td><strong>Renren</strong></td>
<td>1 year</td>
<td>1M</td>
<td>14M</td>
<td>11M</td>
</tr>
<tr>
<td><strong>(Complete)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Facebook</strong></td>
<td>2 years</td>
<td>49K</td>
<td>339K</td>
<td>64K</td>
</tr>
<tr>
<td><strong>(Regional Network)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>YouTube</strong></td>
<td>5 months</td>
<td>1M</td>
<td>3M</td>
<td>3M</td>
</tr>
<tr>
<td><strong>(Snowball Crawl)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[1] ZHAO, SALA, WILSON, WANG, GAITO, ZHENG, and ZHAO. Multi-scale dynamics in a massive online social network. IMC 2012.
Evaluation Metric

• Divide network into sequential graph snapshots \( (G_1, G_2, \ldots, G_N) \), where \( G_t = <V_t, E_t> \) at time \( T_t \)
  - \( |E_{t+1}| - |E_t| \) is the same
  - \( \sim 20 \) graph snapshots: \( T_{t+1} - T_t \) is within 2 weeks \( \sim 1 \) month

• Given two graph snapshot \( G_{t-1} \) and \( G_t \), predict \( k \) new edges in \( G_t \)
  - \( k = \# \) of new edges formed among \( V_{t-1} \)
  - Precision: \[ \frac{\text{top } k \text{ edges } \cap \text{actual new edges}}{k} \times 100\% \]
Outline

• Motivation

• Methodology

• Results from Empirical Study
  – Metric-based algorithms
  – Classification-based algorithms

• New Technique based on Temporal Dynamics
Metric-based Prediction

• Best precision (%) across snapshots of selected algorithms

<table>
<thead>
<tr>
<th>Network</th>
<th>RA*</th>
<th>Katz</th>
<th>Rescal</th>
<th>AA*</th>
<th>CN*</th>
<th>JC</th>
<th>PPR</th>
<th>SP</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renren</td>
<td>3.52</td>
<td>0.82</td>
<td>0.09</td>
<td>3.22</td>
<td>2.40</td>
<td>1.72</td>
<td>2.44</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Facebook</td>
<td>4.43</td>
<td>9.41</td>
<td>4.45</td>
<td>6.82</td>
<td>6.17</td>
<td>1.21</td>
<td>1.06</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>YouTube</td>
<td>0.44</td>
<td>0.98</td>
<td>1.75</td>
<td>0.53</td>
<td>0.59</td>
<td>0.22</td>
<td>0.23</td>
<td>0.00</td>
<td>0.38</td>
</tr>
</tbody>
</table>

• Key findings
  – No single winner across all networks
  – Some relatively good, e.g., RA (Resource Allocation), Katz, Rescal
  – Some consistently bad, e.g., SP (Shortest Path), PA (Preferential Attachment)

Choosing Metric-based Algorithm

• Q: given a network, can one predict the best link prediction algorithm?

• A: train a multiclass classifier (decision tree)
  – Input features: network properties
    o e.g., # of nodes/edges, node degree distribution (average, standard deviation), clustering coefficient
  – Each snapshot is a data point, total # = 69
Choosing Metric-based Algorithm

- Not a definitive guide, but general trends
  - Network with high heterogeneity in node degree: Rescal
  - Sparse network with low heterogeneity in node degree: Katz
  - Dense network with low heterogeneity in node degree: RA
Sources of Failure

- Certain nodes are more likely to create edges[1]
  - Temporal aspect: nodes recently active in edge creation
    - Captured by idle time = time gap since last edge created
    - Worse algorithms: not focused on recently active nodes
  - Structural aspect: e.g., impact in node degree

[1] ZH AO, SALA, WILSON, WANG, GAITO, ZHENG and ZH AO. Multi-scale dynamics in a massive online social network. IMC 2012
Classification-based Prediction

• SVM*: consistently the best classifier

• Compare precision (%) with top metric-based algorithms

<table>
<thead>
<tr>
<th>Network</th>
<th>SVM</th>
<th>RA**</th>
<th>Katz</th>
<th>Rescal</th>
<th>AA**</th>
<th>CN**</th>
<th>LP</th>
<th>PPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renren</td>
<td>2.76</td>
<td>1.24</td>
<td>0.37</td>
<td>0.08</td>
<td>0.52</td>
<td>0.32</td>
<td>0.16</td>
<td>2.76</td>
</tr>
<tr>
<td>Facebook</td>
<td>3.02</td>
<td>2.37</td>
<td><strong>2.41</strong></td>
<td>1.46</td>
<td>1.97</td>
<td>1.74</td>
<td>1.55</td>
<td>0.5</td>
</tr>
<tr>
<td>YouTube</td>
<td>1.32</td>
<td>0.86</td>
<td>0.95</td>
<td><strong>1.31</strong></td>
<td>1.06</td>
<td>1.15</td>
<td>1.15</td>
<td>0.17</td>
</tr>
</tbody>
</table>

• Key findings
  − SVM consistently $\geq$ best metric-based algorithms
  − Best features based on good metric-based algorithms
    ○ e.g., top 2 for Facebook: Katz & RA, top 1 for YouTube: Rescal

*: Classification-based algorithms are evaluated on datasets of snowball sampling, due to high complexity.
**: Local Naïve Bayes version of these algorithms.
Summary of Empirical Study

• Classification-based vs. Metric-based
  – Classification-based: higher consistency & complexity
  – Metric-based: some can provide reasonable alternatives

• Prediction precision: quite low across the board, for all algorithms on all snapshots across all 3 datasets
  – Even the best methods: 5~6%
  – The best results tend to come from Facebook data, likely because it is much smaller (33 times fewer nodes)

Link Prediction is far from a solved problem!
Outline

• Motivation

• Methodology

• Results from Empirical Study

• New Technique based on Temporal Dynamics
Proposed Approach

• Seek to integrate temporal dynamics
  – Existing algorithms rely on static approaches

• Identify key patterns on network growth & build “temporal filters”

• Temporal filters: prune search space for potential edges
  – Identify and remove those unlikely-to-form
  – Complement to existing algorithms as a preprocessing step
  – Significantly reduce algorithm complexity
Build Temporal Filters

- Investigate key properties affecting link connection process
  - Distributions of certain property between **actual links** and **unlikely links** are quite different
    - Actual links: potential edges formed in the next snapshot
    - Unlikely links: potential edges did not form in the next snapshot
Build Temporal Filters

- Investigate key properties affecting link connection process
  - Distributions of certain property between actual links and unlikely links are quite different
    - Actual links: potential edges formed in the next snapshot
    - Unlikely links: potential edges did not form in the next snapshot
- Remove potential edges when they fail to meet thresholds for identified properties
  - Thresholds: network specific
  - Identified properties: consistent across networks

![CDF of Potential Edges](image)

> 60% unlikely links < 10% actual links

11/14/16 Qingyun Liu
Boosting Prediction Accuracy

- Relative improvement of precision over all algorithms over all networks
  - Temporal filters affect certain algorithms more

<table>
<thead>
<tr>
<th>Network</th>
<th>Metric-based Improvement (Min / Max)</th>
<th>Classification-based Improvement (Min / Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renren</td>
<td>50% / 1390%</td>
<td>80% / 90%</td>
</tr>
<tr>
<td>Facebook</td>
<td>20% / 470%</td>
<td>20% / 50%</td>
</tr>
<tr>
<td>YouTube</td>
<td>10% / 1470%</td>
<td>10% / 120%</td>
</tr>
</tbody>
</table>

11/14/16
Qingyun Liu
Summary

• Link prediction is a difficult problem
  – Even with our temporal filters, best prediction precision ~ 10% (with only structural information)

• Lessons from existing algorithms
  – Best metric-based algorithm: possible alternative for SVM
  – Provide potential guidelines for choosing algorithms

• Provide “temporal filters” that can greatly improve prediction, by leveraging temporal dynamics
Thank You!
Any Questions?