STACY
Strength of Ties Automatic-Classifier over the Years

Michele A. Brandão, Pedro O. S. Vaz de Melo & Mirella M. Moro
Universidade Federal de Minas Gerais - UFMG, Brazil

SBBD@2017
Outline

1. Introduction
2. Related work
3. Background
4. fast-RECAST versus STACY
5. Experiments and results
6. Conclusions and future work
1. Introduction

Context

- Social networks → relationships and interactions among individuals
- Their models and patterns allow to solve different problems
Example

Dengue @ Brazil

Mining over Twitter data

Outbreaks detection

Predict future outbreaks

Plan combat actions properly
● **Social ties**: the strength of ties

● **Tie strength persistence and transformation** → *time*
Time
Relevance

- Relevant knowledge
- Different applications
- Academic context
  - Ranking
  - Name disambiguation
1. How to measure tie strength?
"Tie strength may be measured by a combination of the amount of time, the cooperation intensity and the reciprocal services that characterize the tie."

[Granovetter, 1973]
2. How to analyze tie strength?
1. New algorithm (called STACY) to measure tie strength in large SN in acceptable run-time

2. Tie strength dynamism analyses over co-authorship networks of different areas
   a. Link persistence analyses
   b. Link transformation analyses
2. Related work
Related Work over Tie Strength

Tie strength overview

- Granovetter [1973]
- **Work** ties [Castilho et al., 2017]
- **Friendship** ties [Seo et al., 2017; Zignani et al., 2016]
- **Contact** ties [Wiese et al., 2015]
- **Developer** ties [Alves et al., 2016; Batista et al., 2017]
Temporal networks

- Temporal aspects has not been largely explored yet
- Kostakos [2009] and Nicosia et al. [2013] propose a set of network properties that consider the temporal aspect
A Justification for Our Work

- Most work focus only on strong and weak ties (e.g. Granovetter [1973])
- RECAST (Vaz de Melo et al. [2015]) is slow and do not consider the duration of the interactions
- Other approaches need social network data (e.g. historic of chat messages)
3. Background
Background

Static aggregated graph

Dynamic graph

Background
Background

Tie Strength

- What is a weak tie?
- A strong tie?
4. fast-RECAST versus STACY
fast-RECAST

fast Random rELationship CIASsifier sTRategy

<table>
<thead>
<tr>
<th>Class</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends (strong)</td>
<td><strong>Frequent</strong> interactions and <strong>many</strong> common friends</td>
</tr>
<tr>
<td>Acquaintances (weak)</td>
<td><strong>Infrequent</strong> interactions and <strong>many</strong> common friends</td>
</tr>
<tr>
<td>Bridge</td>
<td><strong>Frequent</strong> interactions and <strong>not many</strong> common friends</td>
</tr>
<tr>
<td>Random</td>
<td><strong>Infrequent</strong> interactions and <strong>not many</strong> common friends</td>
</tr>
</tbody>
</table>
fast-RECAST

- Build more than one random graph at a time
- Compute edge persistence and topological overlap in parallel
- Optimize memory use

**Algorithm 1** Multiprocessing RECAST (fast-RECAST): a parallelized code to classify edges of $G_t$ as random or social - strong, weak or bridge.

**Require:** $p_{rnd} \geq 0$

1. return $\text{class}(i,j) \ \forall (i,j) \in U_t E_t$
2. Construct $G_t^R$ and set $\text{RND}(G_1), ..., \text{RND}(G_t)$ using $\text{T-RND}$ with different processes for each $\text{RND}(G_k)$ (pool.map.async)
3. Get $\bar{F}_{to}(x)$ and $\bar{F}_{per}(x)$ from $G_t^R$ using a two-dimensional size mutable data structure (pandas dataframe)
4. Get $\bar{x}_{to}|\bar{F}_{to}(\bar{x}_{to})$ and $\bar{x}_{per}|\bar{F}_{per}(\bar{x}_{per}) = p_{rnd}$ with a process for each network feature (pool.apply_async)
5. for all edges $(i,j) \in E_t$ do
6. if $\text{per}(i,j) > \bar{x}_{per}$ and $\text{to}(i,j) > \bar{x}_{to}$ then
7. $\text{class}(i,j) \leftarrow \text{Strong}$
8. else if $\text{per}(i,j) > \bar{x}_{per}$ and $\text{to}(i,j) \leq \bar{x}_{to}$ then
9. $\text{class}(i,j) \leftarrow \text{Bridges}$
10. else if $\text{per}(i,j) \leq \bar{x}_{per}$ and $\text{to}(i,j) > \bar{x}_{to}$ then
11. $\text{class}(i,j) \leftarrow \text{Weak}$
12. else
13. $\text{class}(i,j) \leftarrow \text{Random}$
Our new algorithm
Strength of Ties Automatic-Classifier over the Years

- Three features
  - Neighborhood overlap (number of common friends)
  - Edge persistence (interaction frequency)
  - Co-authorship frequency

- Random graphs
Algorithm 2 STACY: a parallelized code to classify weighted edges of $G^W_t$ as eight different tie strength classes.

Input: Weighted temporal aggregated graph - $G^W_t$

Require: $p_{rand} \geq 0$

1: return class(i, j) $\forall (i, j) \in U_t E_t$
2: Construct $G^RW_t$ and set $RND(G^W_1),...,RND(G^W_t)$ using $T$-RND with pool.map_async
3: Get $F_to(x)$ and $F_per(x)$ and $F_coAfrequency(x)$ from $G^RW_t$ using pandas dataframe
4: Get $\bar{x}_{to}/F_to(x_{to})$ and $\bar{x}_{per}/F_per(x_{per})$ and $\bar{x}_{coAfrequency}/F_coAfrequency(x_{coAfrequency})$ $\forall p_{rand}$ with pool.apply_async
5: for all edges $(i, j) \in E_t$ do
6: if $per(i, j) > \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ and $coAfrequency(i, j) > \bar{x}_{coAfrequency}$
7: class(i, j) $\leftarrow$ Class1
8: else if $per(i, j) > \bar{x}_{per}$ and $to(i, j) \leq \bar{x}_{to}$ and $coAfrequency(i, j) > \bar{x}_{coAfrequency}$
9: class(i, j) $\leftarrow$ Class2
10: else if $per(i, j) \leq \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ and $coAfrequency(i, j) > \bar{x}_{coAfrequency}$
11: class(i, j) $\leftarrow$ Class3
12: else if $per(i, j) > \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ and $coAfrequency(i, j) \leq \bar{x}_{coAfrequency}$
13: class(i, j) $\leftarrow$ Class4
14: else if $per(i, j) \leq \bar{x}_{per}$ and $to(i, j) \leq \bar{x}_{to}$ and $coAfrequency(i, j) > \bar{x}_{coAfrequency}$
15: class(i, j) $\leftarrow$ Class5
16: else if $per(i, j) > \bar{x}_{per}$ and $to(i, j) \leq \bar{x}_{to}$ and $coAfrequency(i, j) \leq \bar{x}_{coAfrequency}$
17: class(i, j) $\leftarrow$ Class6
18: else if $per(i, j) \leq \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ and $coAfrequency(i, j) \leq \bar{x}_{coAfrequency}$
19: class(i, j) $\leftarrow$ Class7
20: else
21: class(i, j) $\leftarrow$ Class8
### STACY

**Relationship classes**

<table>
<thead>
<tr>
<th>Class</th>
<th>Interaction frequency</th>
<th>Common Friends</th>
<th>Interaction intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class1 - strong</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Class2 - bridge+</td>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Class3 - transient</td>
<td>low</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Class4 - periodic</td>
<td>high</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Class5 - bursty</td>
<td>low</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Class6 - bridge</td>
<td>high</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Class7 - weak</td>
<td>low</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Class8 - random</td>
<td>low</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>
## 5. Experiments and Results

### Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#nodes</th>
<th>#edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP Articles</td>
<td>837,583</td>
<td>2,935,590</td>
</tr>
<tr>
<td>DBLP Inproceedings</td>
<td>945,297</td>
<td>3,760,247</td>
</tr>
<tr>
<td>PubMed</td>
<td>443,784</td>
<td>5,550,294</td>
</tr>
<tr>
<td>APS</td>
<td>180,718</td>
<td>821,870</td>
</tr>
</tbody>
</table>
fast-RECAST

Relationship classes

DBLP Articles  DBLP Inproceedings  PubMed  APS
Relationship classes

1 - strong
2 - bridge+
3 - transient
4 - periodic
5 - bursty
6 - bridge
7 - weak
8 - random
STACY

Relationship classes

![Bar chart for PubMed showing distribution of relationship classes 1 to 8.](Image)

- **1 - strong**
- **2 - bridge+**
- **3 - transient**
- **4 - periodic**
- **5 - bursty**
- **6 - bridge**
- **7 - weak**
- **8 - random**

![Bar chart for APS showing distribution of relationship classes 1 to 8.](Image)
Relationship classes

- APS - Strong
- APS - Bridge+
- APS - Transient
- APS - Periodic
- APS - Bursty
- APS - Bridge
- APS - Weak
- APS - Random
STACY

Relationship classes

1 - strong
2 - bridge+
3 - transient
4 - periodic
5 - bursty
6 - bridge
7 - weak
8 - random
Temporal Networks

Tie Strength Concept

- Strong ties → persistent
- Weak ties → sporadic
Tie Strength

Varying over Time
80%-20%

- Link persistence
  - STACY classifies strong ties better than fast-RECAST
  - Strong ties and bridges persist more than others

<table>
<thead>
<tr>
<th>Dataset</th>
<th>fast-RECAST</th>
<th>STACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP Articles</td>
<td>0.21</td>
<td>0.39</td>
</tr>
<tr>
<td>DBLP Inproceedings</td>
<td>0.14</td>
<td>0.26</td>
</tr>
<tr>
<td>PubMed</td>
<td>0.21</td>
<td>0.37</td>
</tr>
<tr>
<td>APS</td>
<td>0.67</td>
<td>0.18</td>
</tr>
</tbody>
</table>
## Tie Strength

### Varying over Time

50%-50% - PubMed

<table>
<thead>
<tr>
<th>Class</th>
<th>Periodic</th>
<th>Bridge</th>
<th>Disappear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>0.14</td>
<td>0.12</td>
<td>0.74</td>
</tr>
<tr>
<td>Bridge+</td>
<td>0.05</td>
<td>0.03</td>
<td>0.91</td>
</tr>
<tr>
<td>Transient</td>
<td>0.19</td>
<td>0.06</td>
<td>0.74</td>
</tr>
<tr>
<td>Periodic</td>
<td>0.2</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Bursty</td>
<td>0.05</td>
<td>0.03</td>
<td>0.9</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.09</td>
<td>0.05</td>
<td>0.86</td>
</tr>
<tr>
<td>Weak</td>
<td>0.22</td>
<td>0.1</td>
<td>0.67</td>
</tr>
<tr>
<td>Random</td>
<td>0.07</td>
<td>0.04</td>
<td>0.88</td>
</tr>
</tbody>
</table>

- **Link transformation**
  - Most ties tend to disappear
  - STACY reveals ties tend to change to class 4 (periodic) and class 6 (bridge)

The paper provides more details
6. Conclusions and Future Work
Main Conclusion

STACY can be applied to

- Automatically detect relationship classes
- Differentiate random from social relationships
- Detecting strong ties that persist more
Future Work

- Expanding the study to other collaboration social networks
- Using qualitative research to evaluate tie strength
- Adding other social network features to STACY
Ongoing Work

Link Prediction

<table>
<thead>
<tr>
<th>Class</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class1 - strong</td>
<td>0.47</td>
</tr>
<tr>
<td>Class2 - bridge+</td>
<td>0.14</td>
</tr>
<tr>
<td>Class3 - transient</td>
<td>0.11</td>
</tr>
<tr>
<td>Class4 - periodic</td>
<td>0.23</td>
</tr>
<tr>
<td>Class5 - bursty</td>
<td>0.08</td>
</tr>
<tr>
<td>Class6 - bridge</td>
<td>0.13</td>
</tr>
<tr>
<td>Class7 - weak</td>
<td>0.09</td>
</tr>
<tr>
<td>Class8 - random</td>
<td>0.05</td>
</tr>
</tbody>
</table>
STACY
Strength of Ties Automatic-Classifier over the Years

Michele A. Brandão, Pedro O. S. Vaz de Melo & Mirella M. Moro
{micheleabrandao, olmo, mirella}@dcc.ufmg.br