Relationship Classification in Large Scale Online Social Networks and Its Impact on Information Propagation

Shaojie Tang: IIT
Jing Yuan: Nanjing University
Xuifei Mao: BUPT
Xiang-Yang Li: IIT
Wei Chen: Microsoft Research Asia
Guojun Dai: Hangzhou Dianzi
Outline

• Introduction
• Preliminaries
• Relationship Classification
• Marketing Strategies for Product Information Propagation
• Experiments
Introduction

• Motivation
  – By far, the biggest mistake may be focusing on the quantity of connections, not the quality

• Contributions
  – Relationship classification problem
  – Information propagation problem
  – Experiments
Preliminaries

• Social network model
  – $G=(V, E)$
  
  
  \[ \mathcal{L} = \{ \ell_0, \ell_1, \ell_2, \cdots, \ell_k \} \]

  \(\text{link } e \in E \text{ has one of multiple labels } \ell(e) \in \mathcal{L}\)

• Problems
  
  – Relationship classification
    • Assign each link $e$ a label (b.o. K), and MAX the accuracy
  
  – Maximize product information propagation
    • Find a set of initial active users under the budget constraint s.t. the total weight of the users they can affect is maximized
Relationship Classification

• Two properties of OSN
  – Network transitivity ([0.1, 0.5])
  – Community structure (cluster)

• Two observations of Renren net
  – Network transitivity + Community Structure + Relationship
    • >85% for same relationship
    • ~1% for different relationships
Q value: \( Q_S = \frac{N_{inner}}{N_{all}} \)

<table>
<thead>
<tr>
<th>User ID</th>
<th>( Q_H )</th>
<th>( Q_C )</th>
<th>( Q_O )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1103</td>
<td>0.79</td>
<td>0.77</td>
<td>0.71</td>
</tr>
<tr>
<td>1004</td>
<td>0.80</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>1011</td>
<td>0.81</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>199999</td>
<td>0.84</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>200000</td>
<td>0.82</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>20001</td>
<td>0.75</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>46398</td>
<td>0.81</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>46399</td>
<td>0.78</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>46400</td>
<td>0.82</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>
Relationship Classification Algorithm (RCA)

\[ V_i \subseteq N(v) \text{ same relationship } \ell_i \text{ with } v \]

- Random walking based
- RCA Design
  - Transition matrix

\[
M_G(v, u) = \begin{cases} 
1/|N_G(v)| & \text{if } u \text{ is } v's \text{ neighbor in } G \\
0 & \text{otherwise.}
\end{cases}
\]

- Shared information table (SIT):
\[ \mathcal{T} \quad \mathcal{T}(\ell_i, e_j) \]
— $t$-affect of $U$ on vertex $v$: the probability that at least one random walk starting from $U$ can visit $v$ within $t$ steps

*Theorem 1:* Given an online social network $G = (V, E)$, assume $|V| = n$, $U \subseteq V$ and each vertex $v \in U$ is assigned a unique index $D(v)$ from $\{1, \cdots, n\}$, then after $t$ steps, the total $t$-affect caused by $U$ is at least:

$$F_G(U, t) = 1_V \times \left[ I_n - \prod_{i=0}^{t} (1_V - 1_U M_G^i)_{Z} \right]$$

where the $i$-th entry of $F_G(U, t)$ denotes the $t$-affect value on the $i$-th vertex.
Marketing Strategies for Product Information Propagation

\[ OPT = \arg \max_{S \subseteq V : c(S) \leq B} w(S) \]

• [1] KDD’03, expected influenced size \( w(S) \)

**Algorithm 2** Hill-Climbing Algorithm

1: \( S_{[1]} := \arg \max_{v \in V ; c(v) \leq B} \{ w(\{v\}) \} \);
2: \( S_{[2]} := \emptyset \);
3: for \( v \in V \setminus S_{[2]} \) do
4: \( v \leftarrow \arg \max_{v \in V ; c(S_{[2]}) + c(v) \leq B} \left\{ \frac{w(S_{[2]} \cup \{v\}) - w(S_{[2]})}{c(v)} \right\} \);
5: \( S_{[2]} = S_{[2]} \cup v \);
6: \( S := \arg \max_{S \in \{S_{[1]}, S_{[2]}\}} \{ w(S) \} \);
Simulations

• Relationship classification
• Relationship based propagation