The link prediction problem for social networks

Alexandra Chouldechova

STATS 319, February 1, 2011
Motivation

- Recommending new friends in online social networks.

- Suggesting interactions between the members of a company/organization that are external to the hierarchical structure of the organization itself.

- Predicting connections between members of terrorist organizations who have not been directly observed to work together.

- Suggesting collaborations between researchers based on co-authorship.
Statement of the problem

- **Link prediction problem:** Given the links in a social network at time $t$ or during a time interval $I$, we wish to predict the links that will be added to the network during the later time interval from time $t'$ to a some given future time.
Statement of the problem

▶ **Link prediction problem:** Given the links in a social network at time $t$ or during a time interval $I$, we wish to predict the links that will be added to the network during the later time interval from time $t'$ to a some given future time.

▶ **Main approach:** Use measures of network-proximity adapted from graph theory, computer science, and the social sciences to determine which unconnected nodes are ‘close together’ in the topology of the network.
General Notation

- Social network $G = \langle V, E \rangle$ (or $G = \langle A, E \rangle$ if nodes are authors)
- An edge $e = \langle u, v \rangle \in E$ represents an interaction between $u$ and $v$ that took place at time $t(e)$
General Notation

- Social network $G = \langle V, E \rangle$ (or $G = \langle A, E \rangle$ if nodes are authors)
- An edge $e = \langle u, v \rangle \in E$ represents an interaction between $u$ and $v$ that took place at time $t(e)$
- For $t < t'$, let $G[t, t']$ denote the subgraph of $G$ consisting of all edges that took place between $t$ and $t'$
General Notation

▸ Social network $G = \langle V, E \rangle$ (or $G = \langle A, E \rangle$ if nodes are authors)

▸ An edge $e = \langle u, v \rangle \in E$ represents an interaction between $u$ and $v$ that took place at time $t(e)$

▸ For $t < t'$, let $G[t, t']$ denote the subgraph of $G$ consisting of all edges that took place between $t$ and $t'$

▸ For $t_0 < t'_0 < t_1 < t'_1$, given $G[t_0, t'_0]$, we wish to output a list of edges not in $G[t_0, t'_0]$ that are predicted to appear in $G[t_1, t'_1]$
General Notation

- Social network $G = \langle V, E \rangle$ (or $G = \langle A, E \rangle$ if nodes are authors)
- An edge $e = \langle u, v \rangle \in E$ represents an interaction between $u$ and $v$ that took place at time $t(e)$
- For $t < t'$, let $G[t, t']$ denote the subgraph of $G$ consisting of all edges that took place between $t$ and $t'$
- For $t_0 < t'_0 < t_1 < t'_1$, given $G[t_0, t'_0]$, we wish to output a list of edges not in $G[t_0, t'_0]$ that are predicted to appear in $G[t_1, t'_1]$
- Let Core $\subset V$ denote the set of all nodes that are incident to at least $\kappa_{\text{training}}$ edges in $G[t_0, t'_0]$ and at least $\kappa_{\text{test}}$ edges in $G[t_1, t'_1]$
Notation for arXiv physics co-authorship network

- $[t_0, t'_0]$ are the three years 1993 – 1996
- $[t_1, t'_1]$ are the three years 1997 – 1999
Notation for arXiv physics co-authorship network

- $[t_0, t'_0]$ are the three years 1993 – 1996
- $[t_1, t'_1]$ are the three years 1997 – 1999
- $G[1993, 1996] = G_{collab} = \langle A, E_{old} \rangle$
Notation for arXiv physics co-authorship network

- \([t_0, t'_0]\) are the three years 1993 – 1996
- \([t_1, t'_1]\) are the three years 1997 – 1999
- \(G[1993, 1996] = G_{collab} = \langle A, E_{old} \rangle\)
- \(E_{new}\) is the set of edges \(\langle u, v \rangle\) such that authors \(u\) and \(v\) co-authored an article sometime during 1997 – 1999 but not during 1993 – 1996
Notation for arXiv physics co-authorship network

- $[t_0, t'_0]$ are the three years 1993 – 1996
- $[t_1, t'_1]$ are the three years 1997 – 1999
- $G[1993, 1996] = G_{collab} = \langle A, E_{old}\rangle$
- $E_{new}$ is the set of edges $\langle u, v \rangle$ such that authors $u$ and $v$ co-authored an article sometime during 1997 – 1999 but not during 1993 – 1996
- Each link predictor $p$ outputs a ranked list $L_p$ of pairs in $A \times A - E_{old}$. List is ordered according to decreasing values of $\text{score}(x, y)$ for $\langle x, y \rangle \in A \times A - E_{old}$
Notation for arXiv physics co-authorship network

- $[t_0, t'_0]$ are the three years $1993 - 1996$
- $[t_1, t'_1]$ are the three years $1997 - 1999$
- $G[1993, 1996] = G_{collab} = \langle A, E_{old} \rangle$
- $E_{new}$ is the set of edges $\langle u, v \rangle$ such that authors $u$ and $v$ co-authored an article sometime during $1997 - 1999$ but not during $1993 - 1996$
- Each link predictor $p$ outputs a ranked list $L_p$ of pairs in $A \times A - E_{old}$. List is ordered according to decreasing values of $\text{score}(x, y)$ for $\langle x, y \rangle \in A \times A - E_{old}$
- $E^*_{new} = E_{new} \cap (\text{Core} \times \text{Core})$, $n = |E^*_{new}|$
Notation for arXiv physics co-authorship network

- $[t_0, t'_0]$ are the three years 1993 – 1996
- $[t_1, t'_1]$ are the three years 1997 – 1999
- $G[1993, 1996] = G_{collab} = \langle A, E_{old} \rangle$
- $E_{new}$ is the set of edges $\langle u, v \rangle$ such that authors $u$ and $v$ co-authored an article sometime during 1997 – 1999 but not during 1993 – 1996
- Each link predictor $p$ outputs a ranked list $L_p$ of pairs in $A \times A - E_{old}$. List is ordered according to decreasing values of $\text{score}(x, y)$ for $\langle x, y \rangle \in A \times A - E_{old}$
- $E^*_{new} = E_{new} \cap (\text{Core} \times \text{Core})$, $n = |E^*_{new}|$
- Evaluate method by taking the top $n$ edge predictions from $L_p$ that are in $\text{Core} \times \text{Core}$ and computing the size of the intersection with $E^*_{new}$
Methods for Link Prediction: Shortest-path

▶ **Shortest-path:** For \( \langle x, y \rangle \in A \times A - E_{old} \), define,

\[
score(x, y) = \text{(negated) length of shortest path between } x \text{ and } y
\]

▶ If there are more than \( n \) pairs of nodes tied for the shortest path length, order them at random.
Methods for Link Prediction: Neighbourhood-based

- Let $\Gamma(x)$ denote the set of neighbours of $x$ in $G_{\text{collab}}$
- **Common neighbours:** Based on the idea that links are formed between nodes who share many common neighbours

$$\text{score}(x, y) = |\Gamma(x) \cap \Gamma(y)|$$
Methods for Link Prediction: Neighbourhood-based

- Let $\Gamma(x)$ denote the set of neighbours of $x$ in $G_{colab}$

- **Common neighbours:** Based on the idea that links are formed between nodes who share many common neighbours

  $$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

- **Jaccard’s coefficient:** Measure how likely a neighbour of $x$ is to be a neighbour of $y$ and vice versa

  $$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$
Methods for Link Prediction: Neighbourhood-based

- Let $\Gamma(x)$ denote the set of neighbours of $x$ in $G_{collab}$
- **Common neighbours:** Based on the idea that links are formed between nodes who share many common neighbours
  \[
  \text{score}(x, y) = |\Gamma(x) \cap \Gamma(y)|
  \]

- **Jaccard's coefficient:** Measure how likely a neighbour of $x$ is to be a neighbour of $y$ and vice versa
  \[
  \text{score}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}
  \]

- **Adamic/Adar:** Assigns large weight to common neighbours $z$ of $x$ and $y$ which themselves have few neighbours $|\Gamma(z)|$
  \[
  \text{score}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}
  \]
Methods for Link Prediction: Preferential attachment

- Based on the premise that a new edge has node \( x \) as its endpoint is proportional to \( |\Gamma(x)| \). i.e., nodes like to form ties with ‘popular’ nodes.

- **Preferential attachment**: Researchers found empirical evidence to suggest that co-authorship is correlated with the product of the neighbourhood sizes.

\[
\text{score}(x, y) = |\Gamma(x)||\Gamma(y)|
\]
Methods for Link Prediction: Ensembles of all Paths

- **Katz_β measure**: Sums over all possible paths between x and y, giving higher weight to shorter paths.

\[
\text{score}(x, y) = \sum_{l=1}^{\infty} \beta^l |\text{paths}_{x,y}^{(l)}|
\]

where \(\beta > 0\) and \(\text{paths}_{x,y}^{(l)}\) is the set of all length-\(l\) paths from \(x\) to \(y\).
Katz $\beta$ measure: Sums over all possible paths between $x$ and $y$, giving higher weight to shorter paths.

$$\text{score}(x, y) = \sum_{l=1}^{\infty} \beta^l |\text{paths}_{x,y}^{(l)}|$$

where $\beta > 0$ and $\text{paths}_{x,y}^{(l)}$ is the set of all length-$l$ paths from $x$ to $y$.

Two variants of the Katz measure are considered

(a) unweighted: $\text{paths}_{x,y}^{(l)} = 1$ if $x$ and $y$ have collaborated and 0 otherwise

(b) weighted: $\text{paths}_{x,y}^{(l)}$ is the number of times that $x$ and $y$ have collaborated.
Methods for Link Prediction: Hitting and Commute times

- Consider a random walk on $G_{collab}$ which starts at $x$ and iteratively moves to a neighbour of $x$ chosen uniformly at random from $\Gamma(x)$.

- The **Hitting Time** $H_{x,y}$ from $x$ to $y$ is the expected number of steps it takes for the RW starting at $x$ to reach $y$.

\[
\text{score}(x, y) = -H_{x,y}
\]
Consider a random walk on $G_{collab}$ which starts at $x$ and iteratively moves to a neighbour of $x$ chosen uniformly at random from $\Gamma(x)$.

The **Hitting Time** $H_{x,y}$ from $x$ to $y$ is the expected number of steps it takes for the RW starting at $x$ to reach $y$.

$$\text{score}(x, y) = -H_{x,y}$$

The **Commute Time** $C_{x,y} = H_{x,y} + H_{y,x}$ is the expected number of steps to travel from $x$ to $y$ then back to $x$.

$$\text{score}(x, y) = -C_{x,y} = -(H_{x,y} + H_{y,x})$$
Methods for Link Prediction: Hitting and Commute times

- Consider a random walk on $G_{collab}$ which starts at $x$ and iteratively moves to a neighbour of $x$ chosen uniformly at random from $\Gamma(x)$.

- The **Hitting Time** $H_{x,y}$ from $x$ to $y$ is the expected number of steps it takes for the RW starting at $x$ to reach $y$.

  $$ score(x, y) = -H_{x,y} $$

- The **Commute Time** $C_{x,y} = H_{x,y} + H_{y,x}$ is the expected number of steps to travel from $x$ to $y$ then back to $x$.

  $$ score(x, y) = -C_{x,y} = -(H_{x,y} + H_{y,x}) $$

- Can also consider stationary-normed versions:

  $$ score(x, y) = -H_{x,y}\pi_y $$

  $$ score(x, y) = -(H_{x,y}\pi_y + H_{y,x}\pi_x) $$
The hitting time and commute time measures are sensitive to parts of the graph far away from $x$ and $y$. 

Rooted PageRank: $\text{score}(x,y) =$ stationary distribution weight of $y$ under this scheme
The hitting time and commute time measures are sensitive to parts of the graph far away from \( x \) and \( y \).

Consider instead the random walk on \( G_{collab} \) that starts at \( x \) that has a probability of \( \alpha \) of returning to \( x \) at each step.

**Rooted PageRank:**

\[
score(x, y) = \text{stationary distribution weight of } y \text{ under this scheme}
\]
Methods for Link Prediction: SimRank

- **SimRank_γ**: Let \( \text{similarity}(x, y) \) be a fixed point of

\[
\text{similarity}(x, y) = \gamma \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{similarity}(a, b)}{|\Gamma(x)||\Gamma(y)|}
\]

where \( \gamma \in [0, 1] \)

| score(x, y) = similarity(x, y) |

- This is the expected value of \( \gamma^\ell \) under the random walk probabilities, where \( \ell \) is the time at which random walks started from \( x \) and \( y \) first meet
Comparisons of the different methods

- Prediction accuracy will be tabulated in terms of relative improvement over a random predictor.
Comparisons of the different methods

- Prediction accuracy will be tabulated in terms of relative improvement over a random predictor.
- The random predictor simply predicts randomly selected pairs of authors from Core who did not collaborate during the training interval 1993 – 1996.
Comparisons of the different methods

- Prediction accuracy will be tabulated in terms of relative improvement over a random predictor.
- The random predictor simply predicts randomly selected pairs of authors from Core who did not collaborate during the training interval 1993 – 1996.
- The probability the random prediction is correct is
  \[
  \frac{1}{\binom{|\text{Core}|}{2} - |E_{old}|}
  \]
- This value ranges from 0.15% in cond-mat to 0.48% in astro-ph.