Data mining --- mining graphs

CAP5771.1
University of South Florida

Xiaoning Qian
Today’s Lecture

1. Complex networks
2. Graph representation for networks
3. Markov chain
4. Viral propagation
5. Google’s PageRank
“New” Science of Networks


The first network/graph problem
- Find a tour crossing every bridge just once, Euler, 1735

*Bridges of Königsberg*
Network Science


- “New” Science
  - Unprecedented number of empirical networks
  - Much larger scale networks
  - Visualization does not convey enough information
  - Computer are much more powerful
  - Highly interdisciplinary
Network Science

- Mining networks/graphs
  - Topology/structure of complex networks
    - Global: degrees, centrality, connectivity, etc.
      - Scale-free (power-law) networks: 6 degree separation?
    - Local: clustering (community), network motifs, etc.
  - Dynamics/behavior of complex networks
    - Global: the topological effect on dynamics
      - How information, virus, disease, rumors, etc. propagate?
    - Local: how individual nodes behave
Complex Networks (Yeast signaling)
Complex Networks (food web)
Complex Networks (friendship)
Complex Networks (romantic relation)

The Structure of Romantic and Sexual Relations at "Jefferson High School"

Each circle represents a student and lines connecting students represent romantic relations occurring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e., we found 63 pairs unconnected to anyone else).
Complex Networks (author citation)
Complex Networks (Internet)
Complex Networks (Web)
Mathematics of Networks (Graphs)

- What is a network/graph?
  - A collection of vertices/nodes joined by edges
  - Different types of vertices and edges:
    - Directed vs. Undirected
    - Weighed vs. Binary
    - Labeled vs. Nonlabeled
    - Bipartite graphs
    - Hypergraphs
  - Mathematically,
    \[ G = \{V, E\} \]
Undirected network: \(<v_i, v_j> \in E \Rightarrow <v_j, v_i> \in E\)
Mathematics of Networks (Graphs)

- Adjacency matrix $L$
  - Symmetric for undirected graphs
  - Square matrix for (self-)graphs; rectangular for bipartite graphs
    
    \[ L_{ij} = e_{ij} \text{ if } <v_i, v_j> \in E \]
  - Matrix analysis for graph mining!

- Simple graphs, connected graphs, complete graphs, …
Node degree $c_i$
- The number of edges incident with vertex $v_i$
- Neighbor set
- Input-, output-degrees
- Degree distributions (power-law)
- ...

Trail (distinct edges), path (distinct nodes), cycle, cut, ...
Markov chain

- Sequential data
What is a Markov chain?

- Finite Markov chain -- \((Q, P)\)
  - \(Q = \{q_1, q_2, \ldots, q_s\}\) : a finite set of states
  - \(P\) : state transition probability matrix
  - Given a sequence of observations: \(x_1x_2\ldots x_n\),
    The probability of the sequence is:
    \[
p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_n|x_1, \ldots, x_{n-1})
    \]
    For first-order time-homogeneous Markov chain:
    \[
p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|x_{i-1}).
    \]
    Hence,
    \[
p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_n|x_{n-1})
    = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1}).
    \]
What is a Markov chain?

- Finite Markov chain -- \((Q, P)\)
  - \(Q = \{B, q_1, q_2, \ldots, q_s\}\) : a finite set of states
  - \(P\) : state transition probability matrix

\[
a_{q_uq_v} = p(x_i = q_v | x_{i-1} = q_u)
\]

initial state probability:

\[
a_{Bq_u} = p(x_1 = q_u) = p(x_1 = q_u | B)
\]

The probability of a sequence can be expressed with \(P\):

\[
p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}).
\]

- Note: The output are states at each time -- states are observable!!
3-state Markov chain model for the weather:

\[ Q = \{\text{Rain (or snow)}, \text{Cloudy}, \text{Sunny}\}; \]

\( P \) is given in the figure;

Initial state probability
Chapman-Kolmogorov Equation

- Chapman-Kolmogorov equation
  \[ p(x_n) = P^{(n-1)} p(x_1) \]

- Limiting distribution (stationary/steady-state distribution)
  - Irreducibility, Periodicity, Ergocity
    \[ p = P p \]

- How to solve \( p \)?
  - Eigen-decomposition of \( P \)
  - Power method
Random walk on graphs (network diffusion) is a Markov process.
What’s behind Google?

- The algorithm of Google---PageRank
PageRank

- What is an important Webpage?
  - There are many Webpages pointing to it
  - Important Webpage point to more important Webpage
  - Importance diffuses based on links between Webpages

- Vertices: Webpages; Edges: hyperlinks;

**HITS**: JM Kleinberg
PageRank

- Diffusion (Random walk) on Web

\[ p_i = \sum_{j=1}^{N} \left( \frac{L_{ij}}{c_j} \right) p_j \]

- \( p_i \): importance for page \( i \); \( L_{ij} \): link from page \( j \) to \( i \);

- Hence, the problem becomes a Markov chain problem (diffusion process):

\[
\begin{bmatrix}
\vdots \\
p \\
\end{bmatrix} = \begin{bmatrix}
0 & \cdots & 0 \\
L & \ddots & \vdots \\
0 & \cdots & \lambda \\
\end{bmatrix}
\begin{bmatrix}
\vdots \\
p \\
\end{bmatrix} = L \begin{bmatrix}
\vdots \\
p \\
\end{bmatrix}
\]
PageRank

- Diffusion (Random walk with restart) on Web

\[
p_i = (1 - \lambda) + \lambda \sum_{j=1}^{N} \left( \frac{L_{ij}}{c_j} \right) p_j
\]

- \(p_i\): importance for page \(i\); \(L_{ij}\): link from page \(j\) to \(i\);

\[
c_j = \sum_{i=1}^{N} L_{ij}
\]
PageRank

- Diffusion (Random walk with restart) on Web

\[ p_i = (1 - \lambda) + \lambda \sum_{j=1}^{N} \left( \frac{L_{ij}}{c_j} \right) p_j \]

Pseudocount \hspace{1cm} \text{Diffusion factor}

- Matrix form:

\[ p = (1 - \lambda) e + \lambda \cdot LD_c^{-1} p \]
PageRank

- How do we solve this?

\[ p = (1 - \lambda) e + \lambda \cdot LD_c^{-1} p \]

- Note that \( p \) is simply for ranking and the absolute values are not critical! WLOG, we assume \( e^T p = N \)

- Hence, the problem becomes a Markov chain problem (diffusion process):

\[
\begin{align*}
    p &= \left[(1 - \lambda)ee^T/N + \lambda LD_c^{-1}\right]p \\
    &= Ap
\end{align*}
\]
Viral propagation

- How does the virus spread over the network?
- Will it become an “epidemic” outbreak?
- How fast the virus will die out or become “epidemic”?
- How we should design “robust” networks to prevent cascading failures?

Mathematical Epidemiology

- SIR (Susceptible-Infected-Recovered) model

- SIS (Susceptible-Infected-Susceptible) model
  - Catching the disease from Infective neighbors (birth rate): $\beta$
  - Recover rate: $\delta$
  - Epidemic threshold: $\tau$

If $\frac{\beta}{\delta} < \tau$, the viral outbreak dies out quickly, but if $\frac{\beta}{\delta} \geq \tau$, the virus may become endemic.
SIS model

- SIS model is again a Markov process!

Resisted infection

Infected by neighbor

Not cured

Susceptible

Infected

Cured

Sum and Product rules in probability!!

Xiaoning Qian (xqian@cse.usf.edu) 10/13/11  CAP5771.1
SIS model

\[ \zeta_i(t) = \prod_{j:\text{neighbor of } i} (p_j(t-1)(1-\beta) + (1 - p_j(t-1))) \]

\[ = \prod_{j:\text{neighbor of } i} (1 - \beta \ast p_j(t-1)) \]

Sum and Product rules in probability!!
SIS model

- With appropriate approximations, we can derive

\[ p(v_i^t = \text{susceptible}) = p(v_i^{t-1} = \text{susceptible}) \xi_i + p(v_i^{t-1} = \text{infective}) \delta \]

\[ 1 - p(v_i^t) = [1 - p(v_i^{t-1})] \xi_i + p(v_i^{t-1}) \delta \]

and

\[ \mathbf{p}_t = \left[ \beta L + (1 - \delta)I \right] \mathbf{p}_{t-1} \]

Sum and Product rules in probability!!
With appropriate approximations, we can derive:

\[ \mathbf{p}_t = \left[ \beta L + (1-\delta)I \right] \mathbf{p}_{t-1} \]

Eigen-decomposition of the matrix \( S = \left[ \beta L + (1-\delta)I \right] \)

\[ \lambda_{i,S} = 1 - \delta + \beta \lambda_{i,L} \quad \forall i \]

\[ \mathbf{p}_t = \sum_i \lambda_{i,S}^t \mathbf{u}_{i,S} \text{tr}(\mathbf{u}_{i,S}) \mathbf{p}_0 \rightarrow 0 \]

Hence,

\[ 1 - \delta + \beta \lambda_{1,L} < 1 \]
SIS model

- With appropriate approximations, we can derive
  \[ p_t = \left[ \beta L + (1 - \delta)I \right] p_{t-1} \]

- Eigen-decomposition of the matrix \( S = \left[ \beta L + (1 - \delta)I \right] \)

Hence,
\[ 1 - \delta + \beta \lambda_{1,L} < 1 \]

- Epidemic threshold:
\[ \tau = \frac{1}{\lambda_{1,L}} \]
Networks/graphs are everywhere and require new tools to study them efficiently and effectively.

Random walk (Markov chain) on graphs and its extension can be a useful technique to “mine” complex networks/graphs

- PageRank
- Viral propagation

Have you learned anything? :)

I am teaching *Biological Network Analysis*, Spring 2012.