Business Intelligence and Process Modelling

F.W. Takes

http://www.liacs.nl/~ftakes/BIPM

Lecture 5: Network Science for BI
Recap

- **Business Intelligence**: anything that aims at providing actionable information that can be used to support business decision making
  - Business Analysis
  - Business Analytics
    - Visual Analytics
    - Predictive Analytics (data mining)
- **Business Intelligence 2.0 — Network Science for BI**
- Process Modelling (April and May)
Network Science for BI
Part I: network science introduction and theory (today)

Part II: BI applications of network science (next week)

Applications in the area of finance, (virtual) economics, (viral) marketing, ...
Network Science for BI

Economic Networks

Risk Contagion (e.g., Bank run)

Monitoring, Analyzing, and Simulating Contagious Risk

Design

BI Applications

Identification, Prediction, and Recommendation of Influential Individuals

Social Networks

Network Modeling & Analysis

Financial Markets, Banking Systems, Supply Chains, ...

Social Contagion (e.g., Word-of-Mouth)

Online Communities, Social Networking Websites, ...

Influence

Daning Hu, IFI, UZH
Data: facts, measurements or text collected for reference or analysis (Oxford dictionary)

- Popular
  - Unstructured data: “tabular” data
  - **Structured data**: “network” data

- Traditional
  - Unstructured data: data that does not fit a certain data structure (text, a list of numeric measurements)
  - Structured data: data that fits a certain data structure (table, tree, **graph/network**, etc.)
Protein Interaction Network
Data → Network Science

- Data
- Data Analysis
- Data Mining
- Data Science
- Big Data

Network science: analyzing “big” structured data consisting of objects connected via certain relationships, in short: networks

Interest from: mathematics, computer science, physics, biology, public administration, social sciences, ...
Networks

- **Objects/entities/nodes/vertices**
- **Relationships/ties/links/edges**
- **Network/graph**: objects and relationships between objects
- Data attributes are annotations on the nodes and the edges
- **Examples:**
  - Online social networks
  - Scientific citation and collaboration networks
  - Webgraphs
  - Biological networks
  - Communication networks
  - Corporate networks
Context: Big Data

Big Data: The four Vs
Volume, Velocity, Variety and Value

VOLUME
Large amounts of data

VELOCITY
Need to be analysed quickly

VARIETY
Different types of unstructured and structured data

VALUE
Extracting business insights and revenue from data

© World Newsmedia Network 2013
## Data Scientist Salaries in United States

Updated Oct 7, 2012 – Salaries posted anonymously by employees and employers

**Change location**: United States – All Cities

<table>
<thead>
<tr>
<th>National</th>
<th>Median</th>
<th>Low to High</th>
</tr>
</thead>
<tbody>
<tr>
<td>$117,500</td>
<td>$93k</td>
<td>$148k</td>
</tr>
</tbody>
</table>

### 61 Salaries: 1–20 of 45 Job Titles

<table>
<thead>
<tr>
<th>Salaries in USD</th>
<th>Avg. Salary</th>
<th>$50k</th>
<th>$100k</th>
<th>$150k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist – Facebook</td>
<td>$122,257</td>
<td>$110k</td>
<td>$145k</td>
<td></td>
</tr>
<tr>
<td>3 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scientist – IMVU</td>
<td>$120,000</td>
<td>$120k</td>
<td>$120k</td>
<td></td>
</tr>
<tr>
<td>3 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scientist – LinkedIn</td>
<td>$105,349</td>
<td>$101k</td>
<td>$109k</td>
<td></td>
</tr>
<tr>
<td>2 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scientist – StumbleUpon</td>
<td>$112,190</td>
<td>$107k</td>
<td>$117k</td>
<td></td>
</tr>
<tr>
<td>2 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scientist – Apollo Group</td>
<td>$123,050</td>
<td>$110k</td>
<td>$138k</td>
<td></td>
</tr>
<tr>
<td>2 Salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Context: Social Media

Social Media Landscape 2013

FredCavazza.net
Social Media Mining

- Social media **platforms**: Facebook, Twitter, LinkedIn, Reddit, YouTube, Blogger, ...
- Platforms generate enormous amounts of (un)structured data
- **Social Media Mining**: analyzing this data in order to get insight in user(s), trends, usage patterns, the platform itself, ...
  - Sentiment Mining
  - Trend Analysis
  - **Social Network Analysis**
Social Media Analytics

Social Media Analytics

Capture
- Gather data from various sources
- Preprocess the data
- Extract pertinent information from the data

Understand
- Remove noisy data (optional)
- Perform advanced analytics: opinion mining and sentiment analysis, topic modeling, social network analysis, and trend analysis

Present
- Summarize and evaluate the findings from the understand stage
- Present the findings

Formalizing . . .

- Social Network Analysis
Formalizing . . .

- Social Network Analysis
- Social Networks
Formalizing . . .

- Social Network Analysis
- Social Networks
- Networks
Formalizing . . .

- Social Network Analysis
- Social Networks
- Networks
- Graphs
## Notation

### Concept
- **Network (graph)**
- **Objects (nodes/vertices)**
- **Relations (links/edges)**
  - Directed — \( E \subseteq V \times V \)
  - Undirected
- **Number of nodes** — \(|V|\)
- **Number of edges** — \(|E|\)
- We assume no self-edges \((u, u)\) and no parallel edges

### Symbol
- \( G = (V, E) \)
- \( V \)
- \( E \)
- \( n \)
- \( m \)
Notation Example

- **Directed graph** $G = (V, E)$
- Nodes $V = \{u, v, w, x, y, z\}$
- Edges $E = \{(u, v), (w, v), (v, w), (v, x), (x, v), (x, w), (y, v), (v, z)\}$
- Node count $n = 6$
- Link count $m = 8$
**Notation Example**

- **Undirected graph** $G = (V, E)$
- Nodes $V = \{u, v, w, x, y, z\}$
- Edges $E = \{\{u, v\}, \{w, v\}, \{v, x\}, \{x, w\}, \{y, v\}, \{v, z\}\}$
- Node count $n = 6$
- Edge count $m = 6$ (counting undirected edges)
- Or: $m = 12$ (counting (symmetric) directed links)
Types of Graphs

- Directed vs undirected graphs
  - Reciprocity/Symmetry: extend to which directed links are mutual
- Weighted vs. unweighted graphs
  - Unweighted: weight of 1 for computational reasons
  - Signed networks: positive and negative weights
- Labeled (annotated) vs. unlabeled nodes / edges
- One-mode (homogenic) vs. two-mode networks
  - Or: multi-mode (heterogenic) networks
- Node and edge attributes
One-mode node-and-edge-labeled network

Source: http://web.stanford.edu/class/cs224w
Two-mode partial node-labeled network

Projecting networks

Figure: Two-mode network

Figure: Projected one-mode network

Directed **Adjacency Matrix**

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 0 & 1 \\
3 & 1 & 1 & 0 & 1 & 1 & 1 \\
4 & 0 & 0 & 1 & 0 & 0 & 0 \\
5 & 0 & 0 & 1 & 0 & 0 & 0 \\
6 & 0 & 1 & 1 & 0 & 0 & 0
\end{array}
\]

- Directed: \( O(n^2) \) memory
- Weighted graphs: integers in cells

**Figure**: \( n = 6 \) and \( m = 12 \)
**Representation**

- **Undirected Adjacency Matrix**

  $\begin{array}{cccccc}
  1 & 2 & 3 & 4 & 5 \\
  2 & 0 & & & & \\
  3 & 1 & 1 & & & \\
  4 & 0 & 0 & 1 & & \\
  5 & 0 & 0 & 1 & 0 & \\
  6 & 0 & 1 & 1 & 0 & 0 \\
  \end{array}$

- Undirected: $O(\frac{1}{2}n(n - 1))$ memory

- Better, but still many zeros

**Figure:** $n = 6$ and $m = 6$
Representation

- **Adjacency List**
  - 1: 3
  - 2: 3 6
  - 3: 1 2 4 5 6
  - 4: 3
  - 5: 3
  - 6: 2 3

- $O(n + 2m)$ memory

**Figure**: $n = 6$ and $m = 6$
Representation

- **Undirected Adjacency List**
  1: 3
  2: 3 6
  3: 4 5 6
  4:
  5:
  6:

- $O(n+m)$ memory

![Figure: $n = 6$ and $m = 6$](image)
(Undirected) **Edge List**

1 3  
2 3  
2 6  
3 4  
3 5  
3 6  

- Commonly used as an input format  
- $O(2m)$ memory

**Figure**: $n = 6$ and $m = 6$
Toy graph: 6 nodes
Corporate network: 1,500 nodes
Collaboration network: 30,000 nodes
Steam social network: 200,000 nodes

Source: http://forum.gephi.org/viewtopic.php?t=2314
Webgraph: 500,000 nodes

Source: Young Hyun, CAIDA, visualized using Walrus
Facebook: 1,000,000,000 nodes
Size and Memory

- **Hyves** online social network
  - \( n = 8,000,000 \) nodes
  - \( m = 1,000,000,000 \) links
- Assume 4 bytes per int (integer)
- Adjacency Matrix: \( n^2 = 8,000,000^2 = 64 \cdot 10^{12} \) bits \( \sim 8 \text{TB} \)
- Adjacency List: \( n + m = 1,008,000,000 \) ints \( \sim 4 \text{GB} \)
- Edge List: \( 2m = 2,000,000,000 \) ints \( \sim 8 \text{GB} \)
- But “smart” graph compression uses only a few bits(!) per edge
So size matters?

- Corporate network \( n = 1,500 \)
- Collaboration network \( n = 30,000 \), \( m = 200,000 \)
- Steam social network \( n = 200,000 \), \( m = 500,000 \)
- Webgraph \( n = 500,000 \), \( m = 8,000,000 \)
- Hyves \( n = 8,000,000 \), \( m = 1,000,000,000 \)
- Facebook \( n = 1,000,000,000 \), \( m = 120,000,000,000 \)
Graph properties

We have seen:
- From 6 to 1,000,000,000 (1 billion) nodes
- From 8 to 120,000,000,000 (120 billion) edges

Measuring only number of nodes and edges is too simple

Other interesting measures:
- Density
- Degree
- Components
- Clustering coefficient
- Distance
Graph properties

- Measuring only number of nodes and edges is too simple
- Other interesting measures:
  - Density
  - Degree
  - Components
  - Clustering coefficient
  - Distance
Sparse networks

- Maximum number of edges $m_{\text{max}}$
  - $m_{\text{max}} = n(n - 1)$ for directed graphs
  - $m_{\text{max}} = \frac{1}{2}n(n - 1)$ for undirected graphs
- Sparse graph if $m \ll m_{\text{max}}$
- Measure sparseness using **density**: $m/n(n - 1)$
- **HYVES**: $8 \cdot 10^6$ nodes, at most $64 \cdot 10^{12}$ edges.
  But network has “only” $1 \cdot 10^9$ edges, so density $0.0000156$. 
Bitcoin Network

- Bitcoin: digital currency
- Peer-to-peer: no central authority
- Blockchain containing all transactions
- Bitcoin network: nodes are addresses (parts of wallets) and directed links are transactions between addresses
- Sparse: $n = 13,086,528$ nodes and $m = 44,032,115$ links
Bitcoin Addresses (daily)
Bitcoin Transactions (daily)
Silk Road Bitcoin Seizure

Source: reddit.com/r/Bitcoin/comments/1prqpu/what_the_silk_road_bitcoin_seizure_transaction
Graph properties

- Measuring only number of nodes and edges is too simple
- Other interesting measures:
  - Density
  - **Degree**
  - Components
  - Clustering coefficient
  - Distance
Degree

Undirected graphs: degree

Directed graphs
- Indegree
- Outdegree

Degree distribution: frequency of each degree value. Follows a power law distribution with a “fat tail”

Undirected graph:
- $deg(v) = 5$

Directed graph:
- $indeg(v) = 4$
- $outdeg(v) = 3$
Bitcoin Network Indegree Distribution

HyVES Degree Distribution
Paths

Concept
- Path
- Path length
- Simple path: no repeated vertices
- Shortest path: path of minimal length
- **Distance**: length of shortest path

Example

Path: \( p = (u, v, z, v, w, x) \)
- \( |p| - 1 = 5 \)

Simple path: \( p' = (u, v, w, x) \)

Shortest path: \( sp = (u, v, x) \)
- \( d(u, x) = |sp| - 1 = 2 \)
Graph properties

- Measuring only number of nodes and edges is too simple
- Other interesting measures:
  - Density
  - Degree
  - Components
  - Clustering coefficient
  - Distance
Connected Components

- What if $d(a, c) = \infty$? (so, no path between nodes $a$ and $c$)
- Multiple connected components
- **Weakly** connected component: subgraph in which there is a path between any pair of nodes if you ignore link direction
- **Strongly** connected component: subgraph in which there is a directed path between any pair of nodes

Component size

- **Giant component**: contains majority of the nodes (usually over 99%)
- Component size distribution often follows a power law
- Singletons: nodes in a component of size 1

Figure: Component size distribution of *Hyves* network
Graph properties

- Measuring only number of nodes and edges is too simple
- Other interesting measures:
  - Density
  - Degree
  - Components
  - Clustering coefficient
  - Distance
Clustering coefficient

- **Clustering coefficient**: extend to which nodes tend to cluster together (form “triangles” of connections)

- Node clustering coefficient for a node $v \in V$:

$$C(v) = \frac{2 \cdot |\{(u, w) \in E : (u, v) \in E \land (v, w) \in E\}|}{\text{deg}(v) \ast (\text{deg}(v) - 1)}$$
Clustering coefficient

- **Clustering coefficient**: extend to which nodes tend to cluster together (form “triangles” of connections)
- Node clustering coefficient for a node $v \in V$:

$$C(v) = \frac{2 \cdot |\{(u, w) \in E : (u, v) \in E \land (v, w) \in E\}|}{\text{deg}(v) \ast (\text{deg}(v) - 1)}$$

$$C(v) = \frac{2 \cdot \text{edges between neighbors of } v}{\text{maximum number of such edges}}$$
Clustering coefficient

- **Clustering coefficient**: extend to which nodes tend to cluster together (form “triangles” of connections)
- Node clustering coefficient for a node \( v \in V \):

\[
C(v) = \frac{2 \cdot |\{(u, w) \in E : (u, v) \in E \land (v, w) \in E\}|}{\text{deg}(v) \ast (\text{deg}(v) - 1)}
\]

\[
C(v) = \frac{2 \cdot \text{edges between neighbors of } v}{\text{maximum number of such edges}}
\]

- Average graph clustering coefficient for a graph \( G \):

\[
C(G) = \frac{1}{n} \cdot \sum_{v \in V} C(v)
\]
Clustering coefficient

Situation A: \( v \) has a clustering coefficient of 0

Situation B: \( v \) has a clustering coefficient of \( \frac{7}{10} = 0.7 \)

Small world networks: high average clustering coefficient compared to a random graph with the same number of vertices

Graph properties

- Measuring only number of nodes and edges is too simple
- Other interesting measures:
  - Density
  - Degree
  - Components
  - Clustering coefficient
  - Distance
Small world experiment

- Stanley Milgram
- 300 letters from Omaha to Boston
- Basic information about intended recipient and request to forward to best acquaintance
- Letters arrived after on average 5.5 steps

Yahoo Small World Experiment

Your objective:
Get a message to this person in as few steps as possible.
On the next page, you will be asked to select one of your Facebook friends, to whom you will forward the message.
You may only select one friend, so choose carefully.

Here is your assigned Target Person:

<table>
<thead>
<tr>
<th>Age</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>City</td>
<td>Berlin</td>
</tr>
<tr>
<td>State/Region</td>
<td>Germany</td>
</tr>
<tr>
<td>Hometown</td>
<td>Berlin, Germany</td>
</tr>
<tr>
<td>Spouse's Name</td>
<td></td>
</tr>
</tbody>
</table>

Education

<table>
<thead>
<tr>
<th>School Name</th>
<th>Grundschule St. Wolfgang Landshut</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Name</td>
<td>University of Newcastle upon Tyne</td>
</tr>
<tr>
<td>Time Period</td>
<td>1999 - 2002</td>
</tr>
</tbody>
</table>
Small World

- Dense **core** containing many hubs
- **Periphery** with many nodes with a small distance to the core
Airline Network

Distance

- Average distance \( \bar{d} = \frac{1}{n(n-1)} \sum_{v, w \in V} d(v, w) \)
- Normal distribution with a “fat tail”

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Nodes</th>
<th>Links</th>
<th>Average degree</th>
<th>Average distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstroPhys</td>
<td>17,903</td>
<td>396K</td>
<td>21</td>
<td>4.15</td>
</tr>
<tr>
<td>Enron</td>
<td>33,696</td>
<td>362K</td>
<td>10</td>
<td>4.07</td>
</tr>
<tr>
<td>Web</td>
<td>855,802</td>
<td>8.64M</td>
<td>10</td>
<td>6.30</td>
</tr>
<tr>
<td>YouTube</td>
<td>1,134,890</td>
<td>5.98M</td>
<td>5.3</td>
<td>5.32</td>
</tr>
<tr>
<td>Flickr</td>
<td>1,624,992</td>
<td>30.9M</td>
<td>5.3</td>
<td>5.38</td>
</tr>
<tr>
<td>Skitter</td>
<td>1,696,415</td>
<td>22.2M</td>
<td>13</td>
<td>5.08</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,213,236</td>
<td>23.5M</td>
<td>11</td>
<td>4.81</td>
</tr>
<tr>
<td>Orkut</td>
<td>3,072,441</td>
<td>234M</td>
<td>76</td>
<td>4.16</td>
</tr>
<tr>
<td>LiveJournal</td>
<td>5,189,809</td>
<td>97.4M</td>
<td>19</td>
<td>5.48</td>
</tr>
<tr>
<td>Hyves</td>
<td>8,057,981</td>
<td>871M</td>
<td>112</td>
<td>4.75</td>
</tr>
</tbody>
</table>

Distance distribution

(distance distribution, sampled over 100000 node pairs)

$f(x)$: frequency of distance $x$

$x$: distance

Distance distribution of a network, showing the frequency of different distances between node pairs.
Erdös number

- Scientific collaboration network
- Edges between scientists who wrote a paper together
- Erdös number: the distance of a scientist (node) to Erdös
- http://www.ams.org/mathscinet/collaborationDistance.html

Figure: Paul Erdös (1913-1996)
Erdős
Movie Actor Network

Source: http://web.stanford.edu/class/cs224w
Six Degrees of Kevin Bacon

- Collaboration network based on co-starring actors
- Variant of “Six degrees of Separation”
- Edges between actors indicate they played in a movie together
- http://www.thekevinbacongame.com

Figure: Kevin Bacon (1958)
Small world networks

- Fat-tailed power-law degree distribution
- Giant **component**
- Low average node-to-node **distance**
- Sparse networks (low **density**)
- High average **clustering coefficient**

Topics in SNA

- Graph Representation and Structure
- Paths and Distances
- Graph Evolution, Link Prediction
- Spidering and Sampling
- Centrality
- Visualization Algorithms and Tools
- Graph Compression
- Community Detection
- Contagion, Gossiping and Virality
- Privacy, Anonymity and Ethics
Lab session March 5

- Get started with Assignment 2
- Sign up on algorithmia.io
- Create a dummy algorithm on algorithmia and make sure that you know how basic input and output works
- Define and extract customer-based features from the sales data
- Report any questions