Social Data Management
Introduction, Data Models, and Measures

Silviu Maniu

November 19th, 2018

Université Paris-Sud

Table of contents

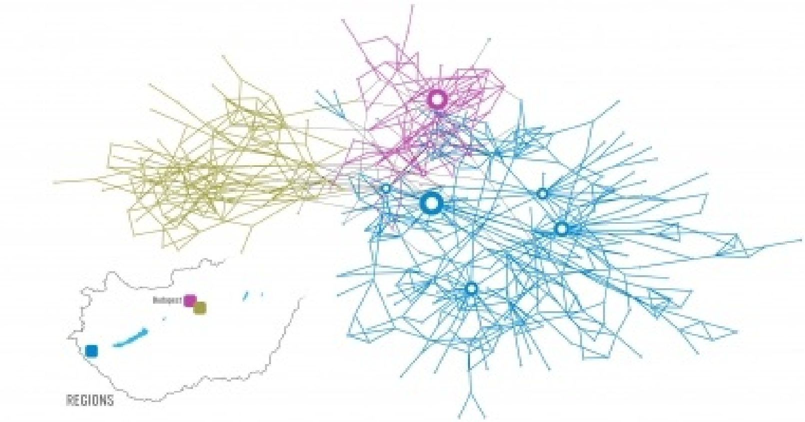
Social Networks

Social networks are an abstract representation of the relationships between *human beings*

They occur in multiple domains (example):

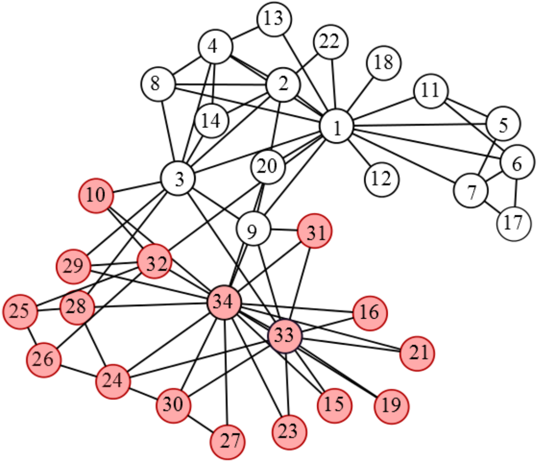
- in an organization, e.g., company, class, ...
- in a professional domain, e.g., physics researchers
- on the Web, e.g., Facebook friends, Twitter followers

Example: Organizations



from A.-L. Barábasi, “Network Science”

Example: Karate Club



by CuneytAkcora, CC BY-SA 4.0 via Wikimedia Commons

Example: Web Social Networks

by Michael Coghlan, CC BY-SA 2.0 via Flickr

Structure of the Course

- we will study the models and measures used for graph analysis
- we will find the properties that distinguish social networks
- we will study some applications of social (graph) data: influence, crowdsourcing, . . .
  众包

Table of contents

Graphs

The most intuitive model for representing social networks are graphs, composed of:

- a set V, representing the *nodes* or *vertices*,
- a *binary relation* E composed of tuples $\{v_1, v_2\} \in V \times V$, representing the *links* or *edges*, and
- optionally, a function $w : E \to$ representing the *weight* of each link.

The resulting graph is represented by the tuple $G = (V, E, w)$. In the following we denote $N = |V|$ and $L = |E|$.

Types of Graphs

Depending on E and w, we can have several types of graphs:

- if $\{v_i, v_j\} \in E$ and $\{v_j, v_i\} \in E$, for any v_i, v_j then the graph is undirected, and directed otherwise,
- if w exists, then the graph is weigthed, and unweighted otherwise.

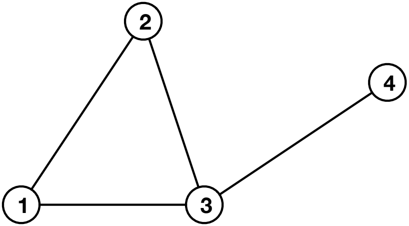
Representing Edges

邻接矩阵和邻接表

Two data structures to represent E:

1. Adjacency Matrix. The adjacency matrix A_G where $a_{ij} = 1$ (or $a_{ij} = w(i, j)$ if weighted graph) for $\{i, j\} \in E$, and $a_{ij} = 0$ otherwise. Good for *dense graphs*, allows random access, but needs $O(V^2)$ space to represent.
2. Adjacency List. The adjacency list $L_G(i)$ is a set of nodes $j \in V$ such that $\{i, j\} \in E$. Good for *sparse graphs*, takes only $O(E)$ space, but no random access. 简朴的

Example: Undirected Graph



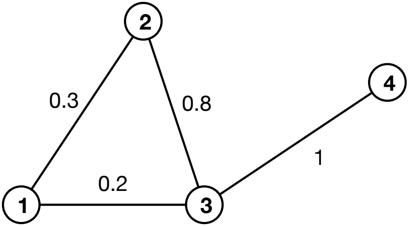
$$V = \{1, 2, 3, 4\}$$

$$E = \{(1,2), (1,3), (2,1), (2,3),$$
$$(3,1), (3,2), (3,4), (4,3)\}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$L(1) = \{2, 3\}$$
$$L(2) = \{1, 3\}$$
$$L(3) = \{1, 2, 4\}$$
$$L(4) = \{3\}$$

Example: Weighted Undirected Graph

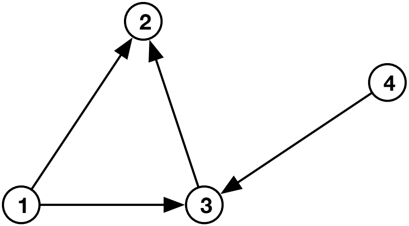


$$
\begin{aligned}
V =& \{1, 2, 3, 4\} \\
E =& \{(1, 2), (1, 3), (2, 1), (2, 3), \\
& (3, 1), (3, 2), (3, 4), (4, 3)\}
\end{aligned}
$$

$$
A = \begin{bmatrix} 0 & 0.3 & 0.2 & 0 \\ 0.3 & 0 & 0.8 & 0 \\ 0.2 & 0.8 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}
$$

$$
\begin{aligned}
L(1) &= \{2, 3\} \\
L(2) &= \{1, 3\} \\
L(3) &= \{1, 2, 4\} \\
L(4) &= \{3\}
\end{aligned}
$$

Example: Directed Graph

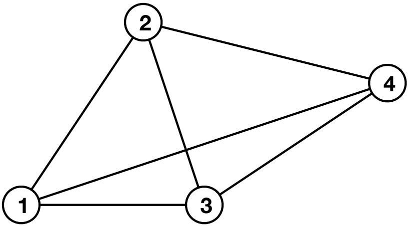


$$V = \{1, 2, 3, 4\}$$
$$E = \{(1,2), (1,3), (3,2), (4,3)\}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$L(1) = \{2, 3\}$$
$$L(2) = \emptyset$$
$$L(3) = \{2\}$$
$$L(4) = \{3\}$$

Example: Complete Graph



$$V = \{1, 2, 3, 4\}$$

$$E = \{(1,2), (1,3), (1,4), (2,1), (2,3), (2,4), (3,1)(3,2), (3,4), (4,1), (4,2), (4,3)\}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$L(1) = \{2, 3, 4\}$$
$$L(2) = \{1, 3, 4\}$$
$$L(3) = \{1, 2, 4\}$$
$$L(4) = \{1, 2, 3\}$$

Table of contents

Degree

The degree $k(i)$ of a node i equals how many other nodes i connects to via links:

$$k(i) = |\{(i,j) \mid j \in V, (i,j) \in E\}|$$

For directed graphs, we have to differentiate between the *incoming* and *outgoing* degree:

$$k^{\text{in}}(i) = |\{(j,i) \mid j \in V, (j,i) \in E\}|$$
$$k^{\text{out}}(i) = |\{(i,j) \mid j \in V, (i,j) \in E\}|$$

Degree Distribution

Denote by p_i the probability that a node has degree i:

$$p_i = \frac{N_i}{N},$$

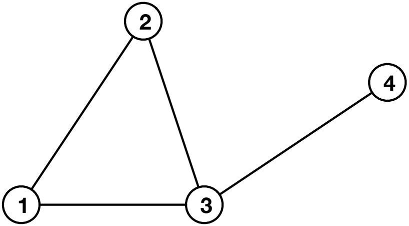
where N_i is the number of nodes of degree i, and N is the total number of nodes in the graph.

This measure defines a distribution:

$$\sum_{i=0}^{\infty} p_i = 1.$$

We can compute the average degree $\langle k \rangle = \sum_{i=0}^{\infty} i \cdot p_i = \frac{L}{N}$.

Example: Degree Distribution



$$
\begin{aligned}
V =& \{1, 2, 3, 4\} \\
E =& \{(1, 2), (1, 3), (2, 1), (2, 3), \\
& (3, 1), (3, 2), (3, 4), (4, 3)\}
\end{aligned}
$$

$$
k(1) = 2,\ k(2) = 2,
$$
$$
k(3) = 3,\ k(4) = 1
$$

$$
\begin{aligned}
p_0 &= 0 \\
p_1 &= 1/4 = 0.25 \\
p_2 &= 2/4 = 0.5 \\
p_3 &= 1/4 = 0.25
\end{aligned}
$$

$$
\begin{aligned}
\langle k \rangle &= 1 \times 0.25 + 2 \times 0.5 + 3 \times 0.25 \\
&= 2
\end{aligned}
$$

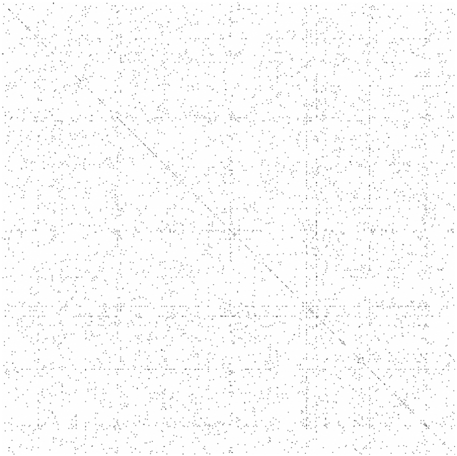
Some Real-World Network Statistics

name	**nodes**	**edges**	$\lvert V\rvert$	$\lvert E\rvert$	$\langle k\rangle$
LiveJournal	users	friendship	4,847,571	68,993,773	14.23
WikiTalk	contributors	communication	2,394,385	5,021,410	2.09
Enron	workers	emails	36,692	183,831	4.99
CondMat	researchers	collaboration	23,133	93,497	4.04
RoadCA	locations	roads	1,965,206	2,766,607	1.40
Web	sites	links	875,713	5,105,039	5.82

More networks and statistics available at `https://snap.stanford.edu/data/`.

Real Networks are Sparse

Our first indication that real networks are different from arbitrary graphs: all the above networks are sparse, with $\langle k \rangle \ll N - 1$.



from Albert-László Barabási, "Network Science"

Paths in Graphs

A path is a sequence of nodes $v_1, v_2, \ldots, v_k$ in V, where each node is a neihbour of the next one.

$$P = \{1, 2, 3, 4\}$$
$$P = \{(1,2), (2,3), (3,4)\}$$

In a *directed* graph, the path can only follow the direction of the arrows.

We can compute the number of paths of length l between two nodes i and j, $N_{ij}^{(l)}$ using the adjacency matrix:

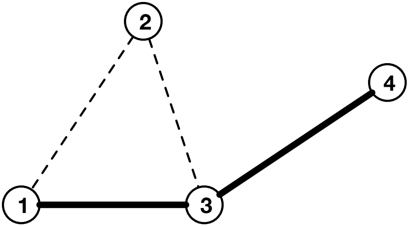
- for $l = 1$ $N_{ij}^{(1)} = A_{ij}$, i.e., the edge between the two nodes,
- otherwise $N_{ij}^{(l)} = [A^l]_{ij}$.

Distances in Graphs

The distance d_{ij} between two nodes i and j in a graphs is:

1. in an *undirected graph*, the number of edges in the shortest path between two nodes, and
2. in a *directed graph*, the weight of the shortest path between two nodes.

Example: Distances



$$V = \{1, 2, 3, 4\}$$
$$E = \{(1,2), (1,3), (2,1), (2,3), (3,1), (3,2), (3,4), (4,3)\}$$

$$d_{14} = 2$$

$$P = (1,3), (3,4)$$

Distances in Graphs

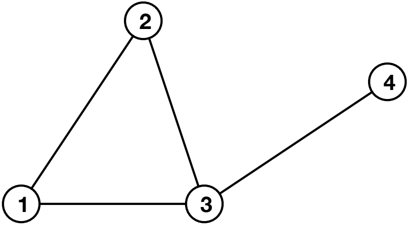
直径

Diameter of a graph d_{max}: the *maximum* distance between any pair of nodes in the graph

Average distance in a graph:

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j} d_{ij}$$

Example: Distances



$$
\begin{aligned}
V =& \{1,2,3,4\} \\
E =& \{(1,2),(1,3),(2,1),(2,3), \\
& (3,1),(3,2),(3,4),(4,3)\}
\end{aligned}
$$

$$d = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix}$$

$$d_{max} = 2$$

$$\langle d \rangle = \frac{16}{12} = 1.33$$

Real Networks Have Low Diameter

For example, Livejournal has a diameter of only 38, despite having several million vertices and edges.

This is known as the six degrees of separation principle – there are not many links separating any two people in the world.

Connectivity

In undirected graphs:

- a connected graph: any two vertices can be joined by a path
- a disconnected graph: made up by two or more connected components

In directed graphs:

- strongly connected if there a path for any vertices i, j in both directions $i \to j$ and $j \to i$.
- weakly connected if there is a path between any vertices i, j *disregarding the direction* of the edges.

Clustering Coefficient

For a node i, the clustering coefficient C_i is the fraction of neighbors that are connected:

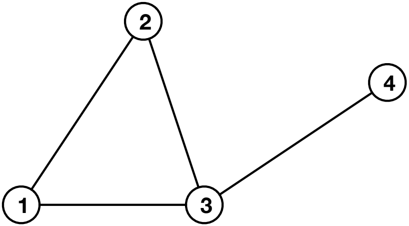
$$C_i = \frac{2e_i \text{ 闭包个数}}{k_i(k_i - 1)},$$

where e_i is the number of between neighbors of i.

The average clustering coefficient is the global measure:

$$\langle C \rangle = \frac{1}{M} \sum_i C_i.$$

Example: Clustering Coefficient



$$
\begin{aligned}
V =& \{1, 2, 3, 4\} \\
E =& \{(1,2), (1,3), (2,1), (2,3), \\
& (3,1), (3,2), (3,4), (4,3)\}
\end{aligned}
$$

$$C_1 = \frac{2 \cdot 1}{2 \cdot 1} = 1$$

$$C_2 = \frac{2 \cdot 1}{2 \cdot 1} = 1$$

$$C_2 = \frac{2 \cdot 1}{3 \cdot 1} = \frac{1}{3}$$

$$C_4 = \frac{2 \cdot 0}{1 \cdot 0} = 0$$

$$\langle C \rangle = \frac{1 + 1 + 1/3}{4} = 0.58$$

Some Real-World Network Statistics

name	**nodes**	**edges**	$\lvert V \rvert$	$\lvert E \rvert$	$\langle C \rangle$
LiveJournal	users	friendship	4,847,571	68,993,773	0.28
WikiTalk	contributors	communication	2,394,385	5,021,410	0.05
Enron	workers	emails	36,692	183,831	0.49
CondMat	researchers	collaboration	23,133	93,497	0.63
RoadCA	locations	roads	1,965,206	2,766,607	0.04
Web	sites	links	875,713	5,105,039	0.51

More networks and statistics available at
`https://snap.stanford.edu/data/`.

Web and Social Networks Have High Clustering Coefficient

Take CondMat: it has a clustering coefficient of 0.63 – intuitively, over 60% of a researcher's collaborators also collaborate between themselves.

Generally, these kinds of networks have a clustering coefficient that is larger than one obtained by chance (more on this later).

Node Centrality Measures

Degree and distances are also part of a class of measures called node centrality measures:

1. vertex centrality is the node's degree k_i
2. closeness centrality is the inverse of the aggregated distances from other nodes $\mathrm{Cl}_i = \frac{1}{\sum_j d_{ji}}$
3. betweennness centrality counts the number of times a nodes is on a shortest path between two nodes
4. eigenvector centrality, e.g., PageRank of a node

Table of contents

Summary

1. We studied some of the important measures in social network analysis: average degree, degree distribution, diameter, and clustering coefficiet.
2. We discovered that they are sparse, with low diameter and high clustering coefficient.
3. Next: *How do these properties emerge in social networks?*

Acknowledgments

The contents is partly inspired by the flow of Chapters 1 and 2 of [Barabási, 2016]. `http://barabasi.com/networksciencebook/`

References i

Barabási, A.-L. (2016).
Network Science.
Cambridge University Press.

Easley, D. and Kleinberg, J. (2010).
Networks, Crowds, and Markets: Reasoning about a Highly Connected World.
Cambridge University Press.

Newman, M. (2010).
Networks: An Introduction.
Oxford University Press.