GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 20: NETWORKS & GRAPHS

YUCHEN ZHAO / DAT-14

I. NETWORKS
II. NETWORK STATICS
III. NETWORK DYNAMICS
IV. LINKEDIN NETWORK MINING

V. FINAL PROJECT WORK SESSION

I. NETWORKS

Q: What is a network?

A: A set of pairwise relationships between objects.

Q: What is a network?

A: A set of pairwise relationships between objects.

The ubiquity of social networks gives rise to many interesting data-oriented questions that can be answered with analytical techniques.

Q: What is a network?

A: A set of pairwise relationships between objects.

The ubiquity of social networks gives rise to many interesting data-oriented questions that can be answered with analytical techniques.

Given a large set of social network data, what types of questions do you think would be interesting to ask?

Some natural questions arise when considering social network data, in particular:

Some natural questions arise when considering social network data, in particular:

- What is the mathematical language for considering network problems?
- What kinds of data structures are well-suited to network analysis?
- What does the network look like?

Some natural questions arise when considering social network data, in particular:

- *What is the mathematical language for considering network problems?*
- *What kinds of data structures are well-suited to network analysis?*
- *What does the network look like?*

These are questions about network representation.

Some natural questions arise when considering social network data, in particular:

- *Who are the most central and/or influential actors in a network?*
- *Can the network be decomposed into coherent smaller groups?*
- *Are any of these groups vital to the functioning of the network?*
- *How does this network compare to other networks?*

Some natural questions arise when considering social network data, in particular:

- *Who are the most central and/or influential actors in a network?*
- *Can the network be decomposed into coherent smaller groups?*
- *Are any of these groups vital to the functioning of the network?*
- *How does this network compare to other networks?*

These are questions about network structure.

Some natural questions arise when considering social network data, in particular:

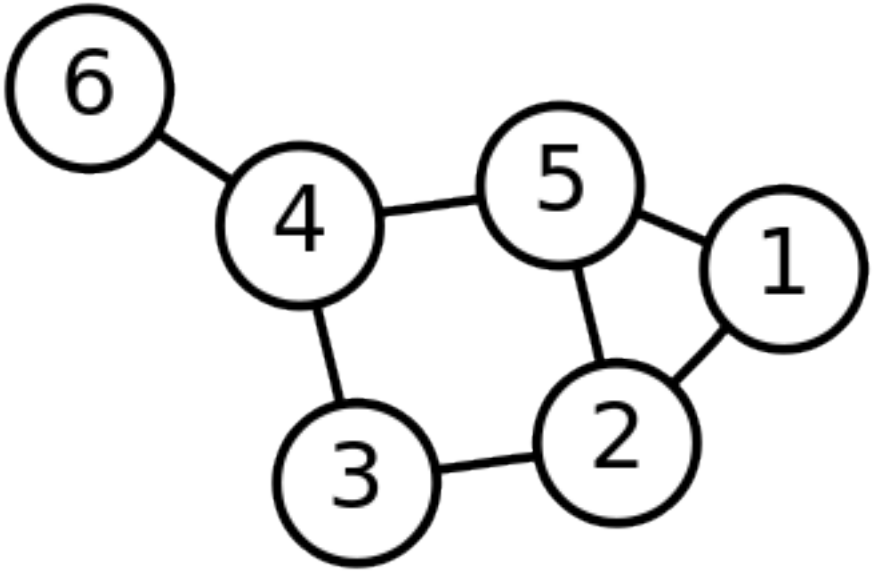
- *How is information propagated through a network?*
- *How does a network acquire or lose members?*
- *How does the structure of the network evolve through time?*
- *How do external events affect the network?*

Some natural questions arise when considering social network data, in particular:

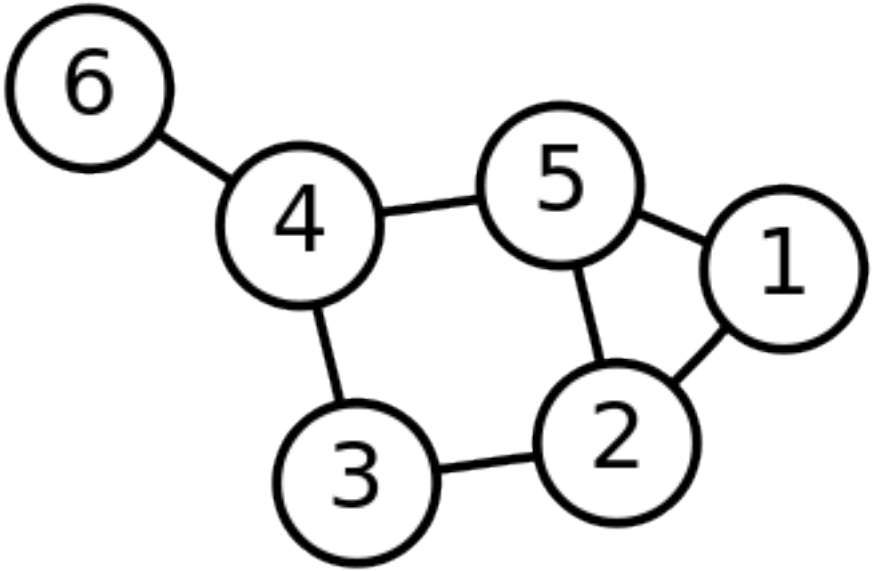
- *How is information propagated through a network?*
- *How does a network acquire or lose members?*
- *How does the structure of the network evolve through time?*
- *How do external events affect the network?*

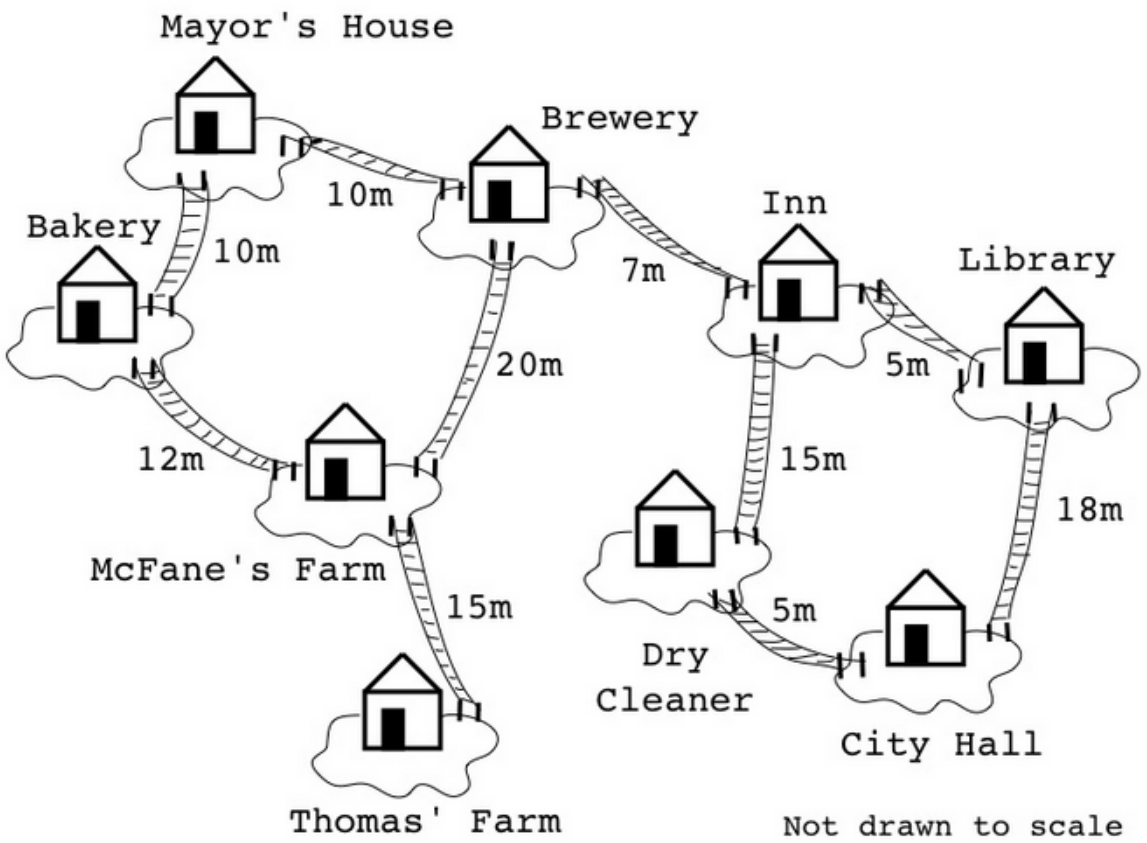
These are questions about network behavior.

The mathematical representation of a network is an object called a **graph**, *which is a configuration of nodes connected by edges.*



Nodes represent actors in the graph, and edges represent the relationships between actors.





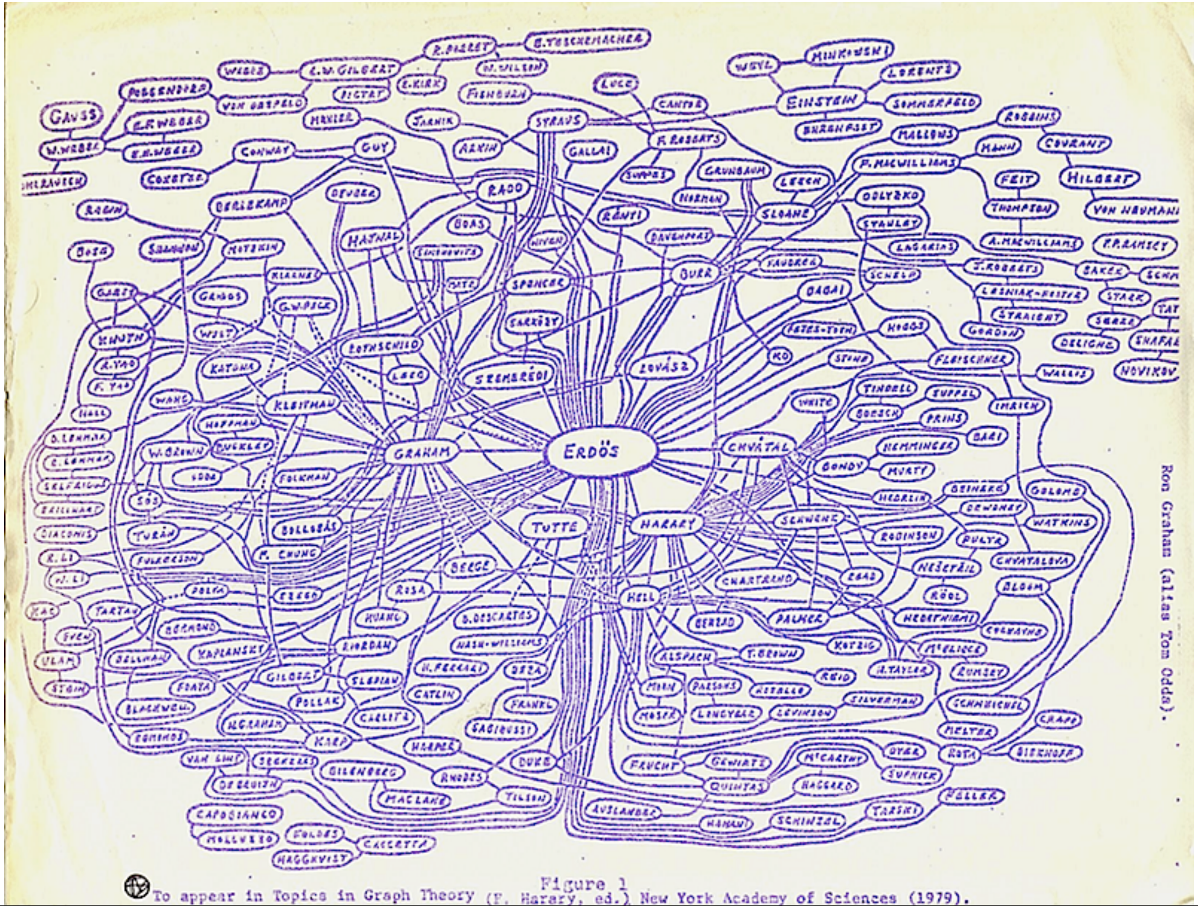
NOTE

A weighted graph contains edges associated with real-valued numbers, eg to measure distance or importance.

NOTE

A directed graph has edges that point from one node to another.

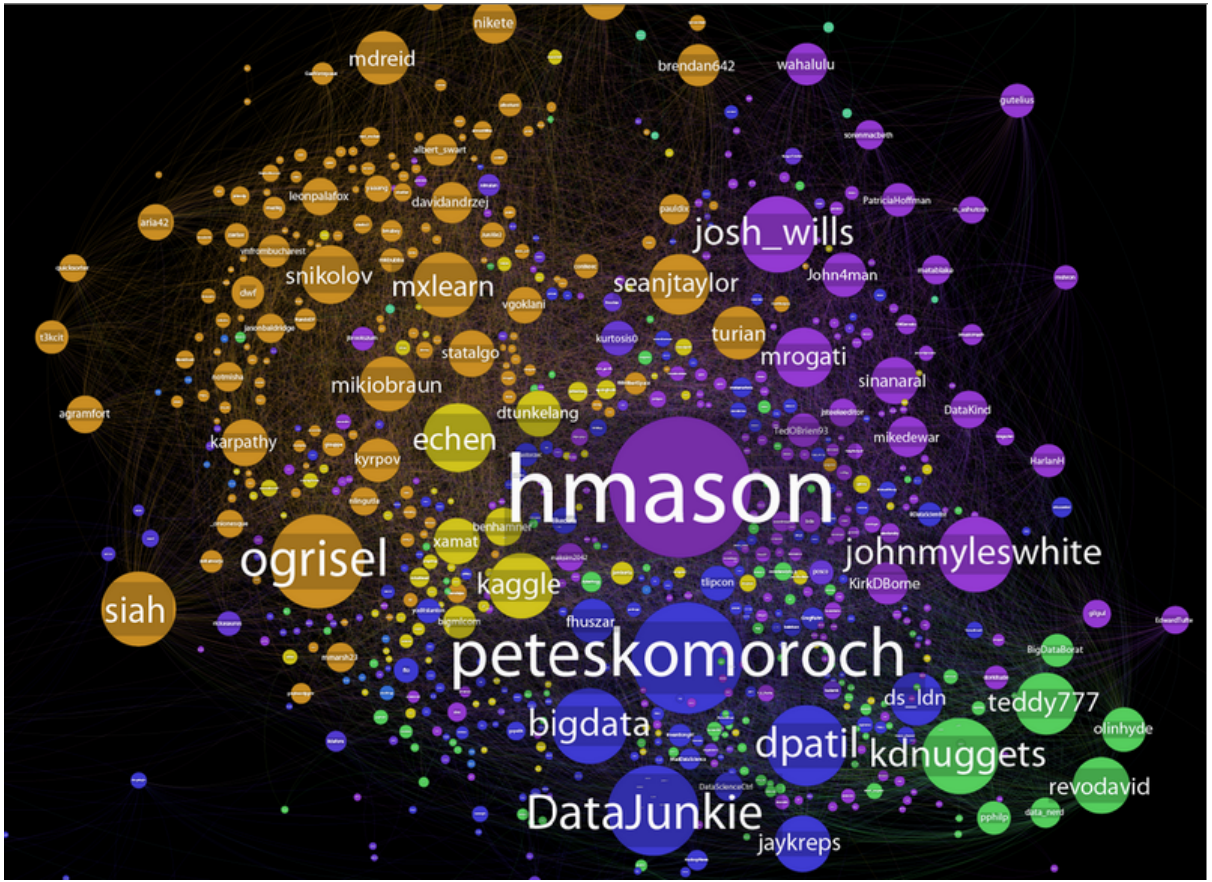
NETWORK REPRESENTATION



NETWORK REPRESENTATION


facebook
December 2010

NETWORK REPRESENTATION



In practical terms, we need some data structures to represent and manipulate our network data.

In practical terms, we need some data structures to represent and manipulate our network data.

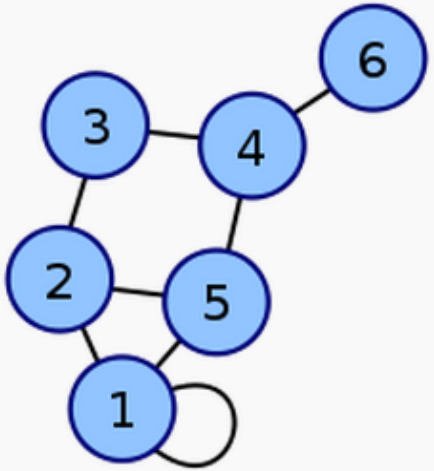
One common graph representation is the **adjacency matrix**.

In practical terms, we need some data structures to represent and manipulate our network data.

Adjacency Matrix:

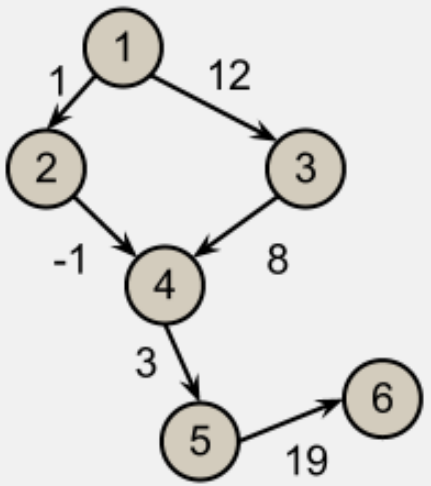
An n*-node undirected graph can be represented by a symmetric* n *x* n *adjacency matrix* A *whose nonzero off-diagonal entries* A_{ij} *represent an edge between nodes* i *and* j.

NETWORK REPRESENTATION

Labeled graph	Adjacency matrix
	$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$ Coordinates are 1-6.

NETWORK REPRESENTATION

Weighted Directed Graph & Adjacency Matrix



Weighted Directed Graph

	1	2	3	4	5	6
1	0	1	12	0	0	0
2	-1	0	0	-1	0	0
3	-12	0	0	8	0	0
4	0	1	-8	0	3	0
5	0	0	0	-3	0	19
6	0	0	0	0	-19	0

Adjacency Matrix

NOTE

A directed graph has an asymmetric adjacency matrix.

Can you see why?

Another useful tool is the **adjacency list** *(actually a dict!):*

```
graph = {'A': ['B', 'C'],
         'B': ['C', 'D'],
         'C': ['D'],
         'D': ['C'],
         'E': ['F'],
         'F': ['C']}
```

Does this adjacency dict represent a directed or undirected graph? How could you generalize this to represent a weighted graph?

II. NETWORK STATICS

One key concept in the study of network structure is centrality. The **centrality** *of a node is a measure of its importance in the network.*

One key concept in the study of network structure is centrality. The **centrality** *of a node is a measure of its importance in the network.*

The simplest centrality measure is the **degree** *of a node, which is simply the number of edges connected to it. Using the adjacency matrix notation for an undirected graph, we can express the degree* k_i *of node* i *as:*

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

Another useful centrality measure is based on the idea of shortest-distance (or geodesic) paths through the graph.

Another useful centrality measure is based on the idea of shortest-distance (or geodesic) paths through the graph.

If σ_{st} *is the number of geodesic paths from node* s *to node* t*, and* $\sigma_{st}(v)$ *is the number of these paths that cross node* v*, then the* **betweenness centrality** *of node* v *is given by:*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Another useful centrality measure is based on the idea of shortest-distance (or geodesic) paths through the graph.

If σ_{st} *is the number of geodesic paths from node* s *to node* t *and* $\sigma_{st}(v)$ *is the number of these paths that cross node* v*, then the* **betweenness centrality** *of node* v *is given by:*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

NOTE

Betweenness centrality measures the proportion of geodesic paths passing through a node.

This gives an idea of the node's influence in the network.

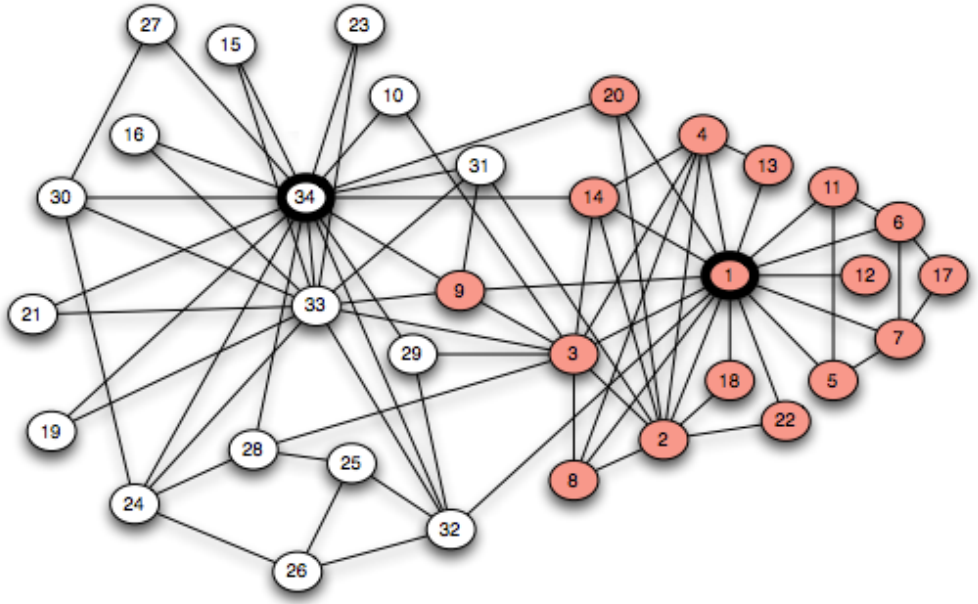


Figure 3.13: A karate club studied by Wayne Zachary [421] — a dispute during the course of the study caused it to split into two clubs. Could the boundaries of the two clubs be predicted from the network structure?

Geodesic paths form the basis of another well-known property of networks called the small-world effect.

Geodesic paths form the basis of another well-known property of networks called the small-world effect.

Specifically, most networks have a mean geodesic distance between nodes that is small compared to the network size as a whole.

Geodesic paths form the basis of another well-known property of networks called the small-world effect.

Specifically, most networks have a mean geodesic distance between nodes that is small compared to the network size as a whole.

A famous study in the 1960s asked participants to try to get a letter to a particular individual by passing it from one acquaintance to another, and found that the mean geodesic distance in this case was about 6.

Geodesic paths form the basis of another well-known property of networks called the small-world effect.

Specifically, most networks have a mean geodesic distance [illegible] *nodes that is small compared to the network size as a whol*

A famous study in the 1960s asked participants to try to get a letter to a particular individual by passing it from one acquaintance to another, and found that the mean geodesic distance in this case was about 6.

NOTE

This is where the phrase “six degrees of separation” comes from.

III. NETWORK DYNAMICS

Suppose we're interested in the idea of how information (or behavior) spreads through a network:

Suppose we're interested in the idea of how information (or behavior) spreads through a network:

- *How do members of a social network influence each other to adopt a new technology/product/behavior?*
- *How did information about the bin Laden raid spread over Twitter?*
- *What's the best way to use a social network to market your product?*

There are two primary methods of influence in social networks:

There are two primary methods of influence in social networks:

informational effects – people observe the decisions of their network neighbors & gain indirect information that lead them to try the innovation themselves

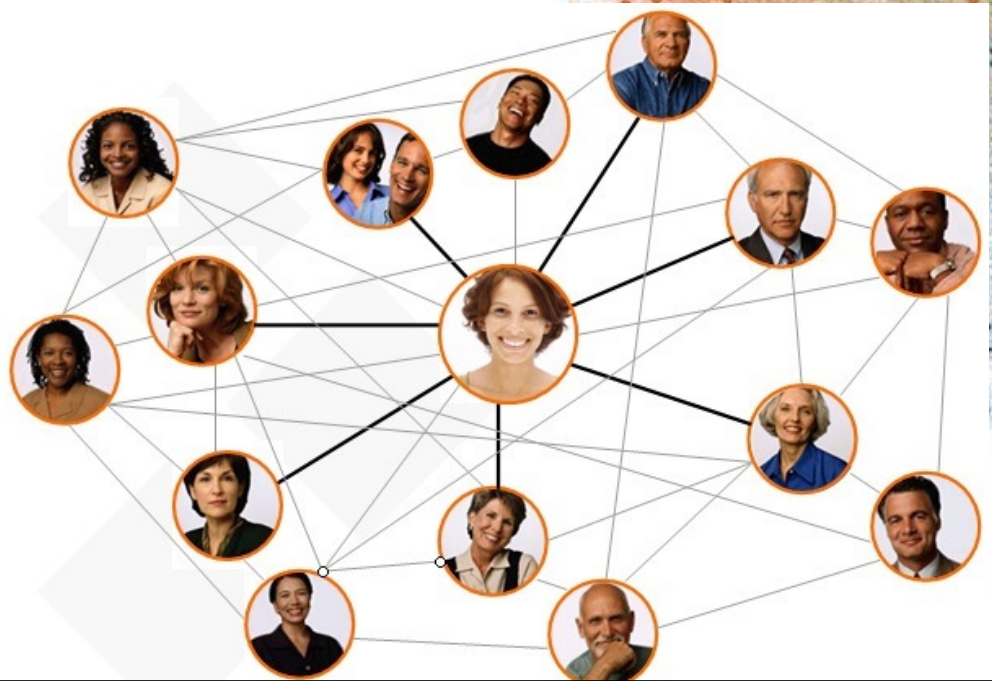
There are two primary methods of influence in social networks:

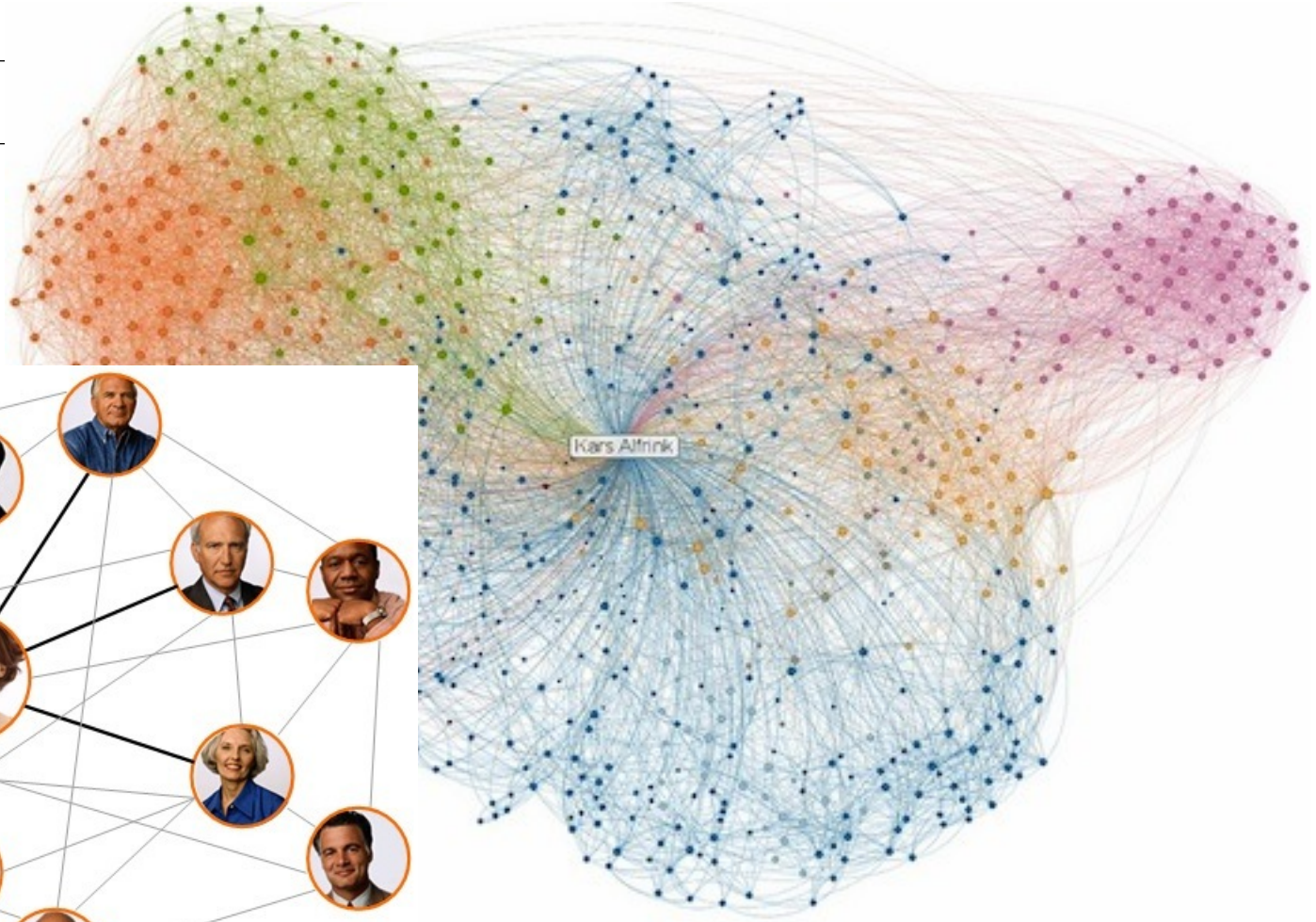
informational effects – people observe the decisions of their network neighbors & gain indirect information that lead them to try the innovation themselves

direct benefit effects – people may have incentives to use the same products/technology/etc as their network neighbors

IV. LINKEDIN NETWORK MINING

LINKEDIN NETWORK





Application 1:

finding out who you really are

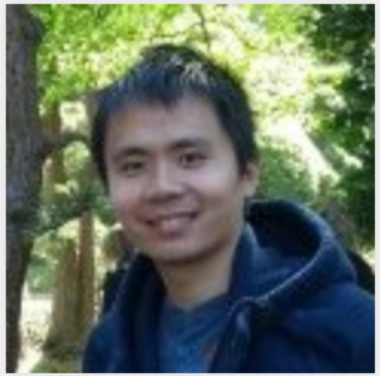
in online social networks

Understanding users' social roles
is crucial to many
social network applications

including
advertising targeting,
marketing,
personalization,
recommendation, etc.

Finding out who you really are...

seems a trivial problem…



Yuchen Zhao 1st

Principal Data Scientist at AppDynamics

San Francisco Bay Area | Internet

Previous General Assembly, Sumo Logic, Scissorsfly

Education University of Illinois at Chicago

Send a message

500+ connections

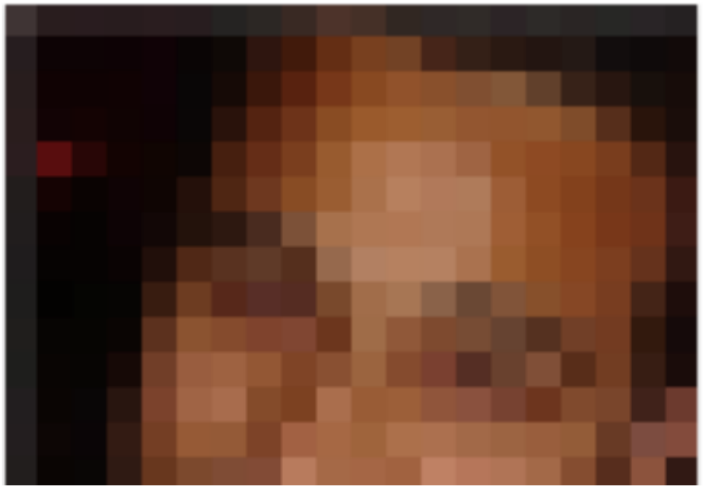
But… really?

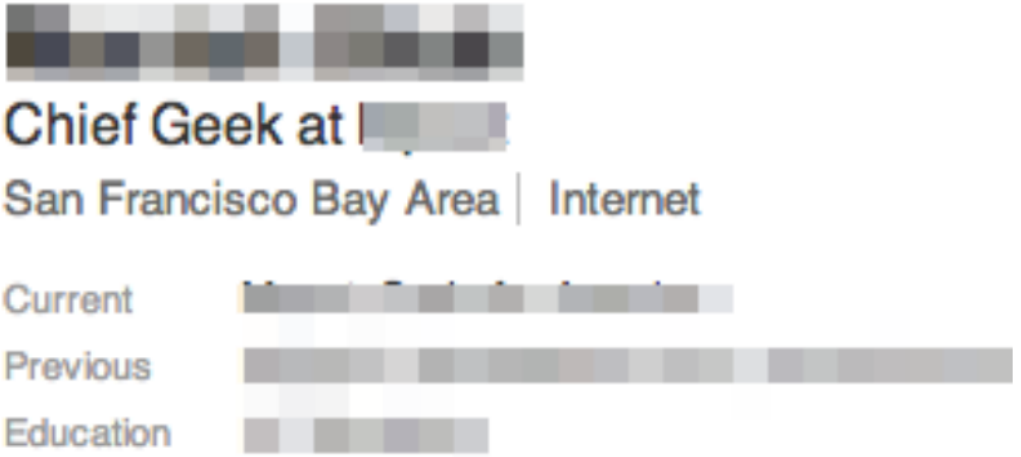
Code Monkey at

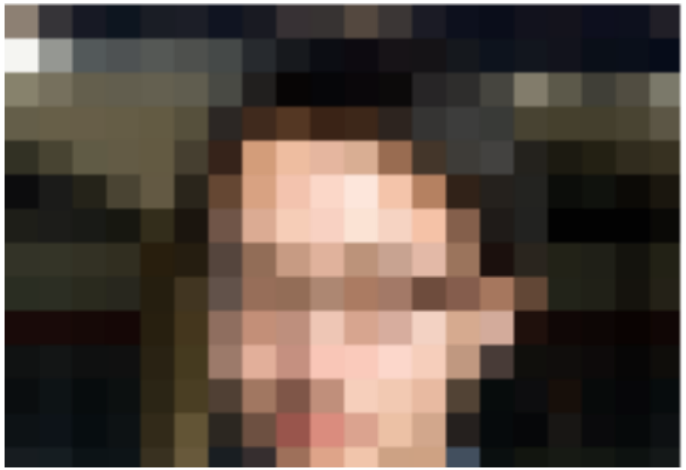
Wilmington, Delaware | Computer Software

Previous

Education




Chief Geek at
San Francisco Bay Area | Internet
Current
Previous
Education



[redacted]

<script>alert('Makes cool stuff');</script>

San Francisco Bay Area | Internet

Current [redacted]

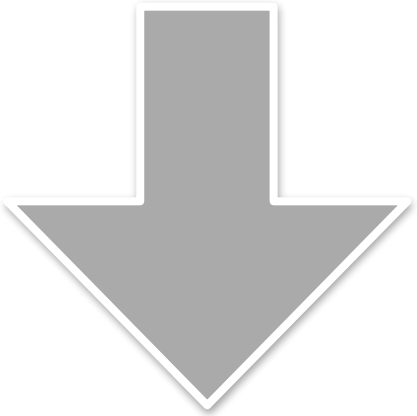
Previous [redacted]

Education [redacted]

also, the data can be:
Missing
&
Outdated

manually labeling is
time-consuming
and
error prone

Human learning



Machine learning

a classification problem!

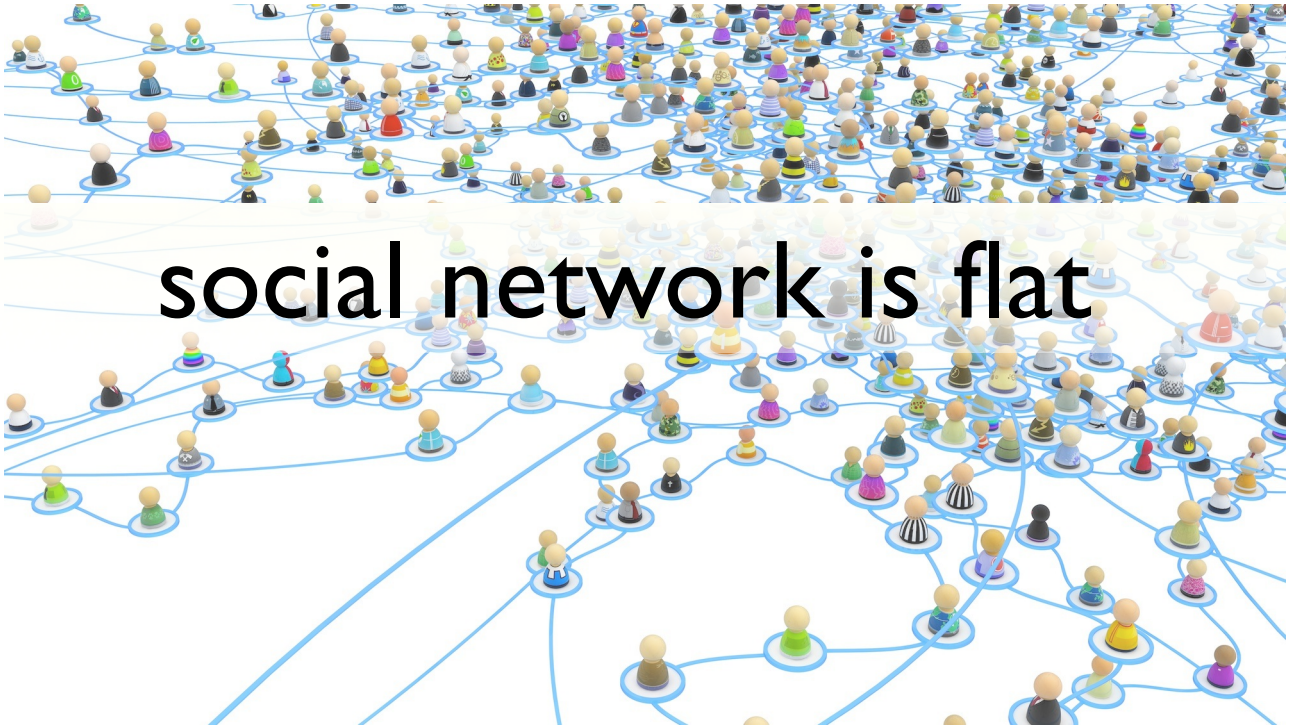
a classification problem
on graph data!

paper discussion:

Inferring Social Roles and Statuses in Social Networks

pdf: http://www.cs.uic.edu/~yzhao/research/papers/fp0200c-zhao.pdf

Application 2:
discovering who is your boss


social network is flat

Can we discover who is who's boss?

A more ambitious question:

Can we infer org charts from a professional social network?

patent discussion:

Techniques for inferring an organizational hierarchy from a social graph

source: https://www.google.com/patents/US20140214945

Application 3:

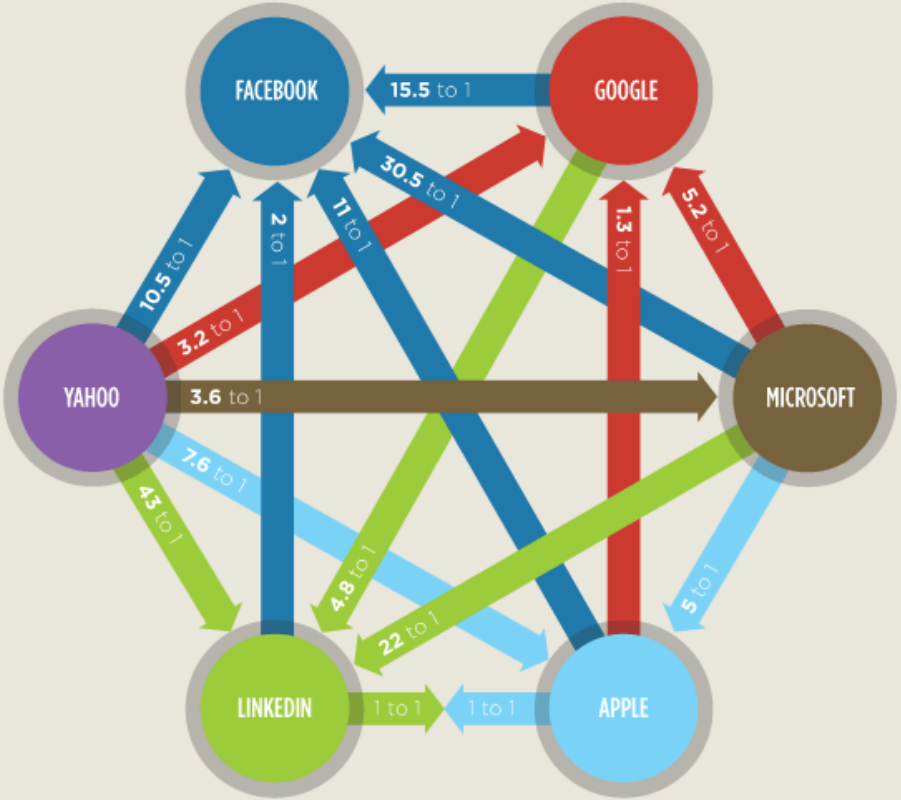
which the next company you should join?

NETWORK MINING

Talent Traffic

topprospect DATA LABS

Ratio is number of employees moving from **Company A** to **Company B** for every one employee going in the other direction. Arrows point to company winning the talent battle.



paper discussion:

Magnet Community Identification on Social Networks

pdf: http://www.cs.uic.edu/~yzhao/research/papers/rt145-wang.pdf