Why is the role of networks in CS, Info science, Social science, (Physics, Economics, Biology) expanding?

- More data
- Rise of the Web and Social media
- Shared vocabulary between (very different fields)
Reasoning about Networks

How do we reason about networks

- **Empirical**: look at large networks and see what you find
- **Mathematical models**: probabilistic, graph theory
- **Algorithms** for analyzing graphs

What do we hope to achieve from models of networks?

- Patterns and statistical properties of network data
- Design principles and models
- Understand why networks are organized the way they are (predict behavior of networked systems)
Why networks?

- Network data is increasingly available:
  - Large on-line computing applications where data can naturally be represented as a network:
    - On-line communities: Facebook (120 million users)
    - Communication: Instant Messenger (~1 billion users)
    - News and Social media: Blogging (250 million users)
  - Also in systems biology, health, medicine, ...
- Network is a set of weakly interacting entities
- Links give added value:
  - Google realized web-pages are connected
  - Collective classification
Introduce properties, models and tools for
- large real-world networks
- diffusion processes in networks through real mining applications

Goal: find patterns, rules, clusters, outliers, ...
- in large static and evolving graphs
- In processes spreading over the networks
Networks: rich data

- Internet (a)
- Citation network (b)
- World Wide Web (c)
- Sexual network (d)
- Dating network (e)
Networks of the Real-world (1)

- Information networks:
  - World Wide Web: hyperlinks
  - Citation networks
  - Blog networks

- Social networks:
  - Organizational networks
  - Communication networks
  - Collaboration networks
  - Sexual networks
  - Collaboration networks

- Technological networks:
  - Power grid
  - Airline, road, river networks
  - Telephone networks
  - Internet
  - Autonomous systems
Networks of the Real-world (2)

- Biological networks:
  - metabolic networks
  - food webs
  - neural networks
  - gene regulatory networks

- Language networks:
  - Semantic networks

- Software networks:
  - Call graphs
  - ...

Yeast protein interactions

Semantic network

Language network

Software network
Mining social networks has a long history in social sciences:

- Wayne Zachary’s PhD work (1970-72): observe social ties and rivalries in a university karate club
- During his observation, conflicts led the group to split
- Split could be explained by a minimum cut in the social network
Traditional obstacle:  
Can only choose 2 of 3:  
- Large-scale  
- Realistic  
- Completely mapped  

Now: large on-line systems leave detailed records of social activity  
- On-line communities: MyScace, Facebook, LiveJournal  
- Email, blogging, electronic markets, instant messaging  
- On-line publications repositories, arXiv, MedLine
Networks: A Matter of Scale

- Network data spans many orders of magnitude:
  - 436-node network of email exchange over 3-months at corporate research lab [Adamic-Adar, SocNets ‘03]
  - 43,553-node network of email exchange over 2 years at a large university [Kossinets-Watts, Science ‘06]
  - 4.4-million-node network of declared friendships on a blogging community [Liben-Nowell et al., PNAS ‘05, Backstrom et al., KDD ‘06]
  - 240-million-node network of all IM communication over a month on Microsoft Instant Messenger [Leskovec-Horvitz, WWW ‘08]
How does massive network data compare to small-scale studies?

Massive network datasets give us more and less:

- **More**: can observe global phenomena that are genuine, but literally invisible at smaller scales
- **Less**: don’t really know what any node or link means. Easy to measure things, hard to pose right questions
- **Goal**: Find the point where the lines of research converge
Networks: Structure & Process

- What have we learned about large networks?
  - **Structure**: Many recurring patterns
    - Scale-free, small-world, locally clustered, bow-tie, hubs and authorities, communities, bipartite cores, network motifs, highly optimized tolerance
  - **Processes and dynamics**:
    - Information propagation, cascades, epidemic thresholds, viral marketing, virus propagation, diffusion of innovation
What is the structure of a large network?
Why and how did it became to have such structure?
Diffusion in Networks

- One of the networks is a spread of a disease, the other one is product recommendations.
- Which is which? 😊
Applications (1)

- Web as a graph:
  - Google PageRank
  - How to estimate webpage importance from the structure of the web-graph?

- Routing in peer-to-peer networks:
  - BitTorrent, ML-donkey, Kazaa, Gnutella
  - Can we find a file in a network without a central server?
Applications (2)

- Marketing and advertising:
  - How to define influence?
  - How to find influencers?
  - Who to give free products to so that we create a network effect?

- Diffusion of information and epidemics:
  - How to trace information as it spreads?
  - How to efficiently detect epidemics and information outbreaks?
Applications (3)

- Friend/link prediction:
  - How to predict/suggest friends in networks?
- Trust and distrust:
  - How to predict who are your friends/foes? Who to trust?
- Community detection:
  - How to find clusters and small communities in social networks
Covers a wide range of network analysis techniques – from basic to state-of-the-art
You will learn about things our heard about:

Six degrees of separation, small-world, page rank, network effects, P2P networks, network evolution, spectral graph theory, virus propagation, link prediction, power-laws, scale free networks, core-periphery, network communities, hubs and authorities, bipartite cores, information cascades, influence maximization, ...

Covers algorithms, theory and applications
It’s going to be fun and hard work 😊
Prerequisites

- Introductory-level background in:
  - Algorithms
  - Graphs
  - Probability
  - Linear algebra
- Programming:
  - To perform data analysis
  - Mostly your choice of language
- Ability to deal with “abstract mathematical concepts”
Course Assistants

- Sonali Aggarwal
  - Office Hours: Thursdays 2:00PM-4:00PM
  - Location: TBD

- Sudarshan Rangarajan
  - Office Hours: Mons & Weds 1:00PM-2:00PM
  - Location: TBD
Text Books

- **Recommended textbook:**

- **Optional books:**
  - Matthew Jackson: Social and Economic Networks
  - Mark Newman: Networks: An introduction
## Work for the course

<table>
<thead>
<tr>
<th>Assignment/Work</th>
<th>Out on</th>
<th>Due on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction Paper</td>
<td>--</td>
<td>October 4</td>
</tr>
<tr>
<td>Assignment #1</td>
<td>September 27</td>
<td>October 11</td>
</tr>
<tr>
<td>Project Proposal</td>
<td>--</td>
<td>October 18</td>
</tr>
<tr>
<td>Assignment #2</td>
<td>October 11</td>
<td>October 25</td>
</tr>
<tr>
<td>Competition</td>
<td>October 25</td>
<td>November 8</td>
</tr>
<tr>
<td>Project Milestone</td>
<td>--</td>
<td>November 15</td>
</tr>
<tr>
<td>Project Poster</td>
<td>--</td>
<td>December 7; 3-6PM</td>
</tr>
<tr>
<td>Presentation</td>
<td>--</td>
<td>December 8; Midnight</td>
</tr>
<tr>
<td>Project Write-up</td>
<td>--</td>
<td></td>
</tr>
</tbody>
</table>
Grading

- **2 homeworks and a competition (30%)**
  - First one out Sept 27
    - Start early, start early, start early, start early, start early
  - **Reaction paper (10%)**
    - A very good way to explore a potential project topic
- **Final project (60%)**
  - Proposal, milestone, final report
  - Poster session
  - Done in groups of 2-3 people
We will have 2 groups of 2-3 students per class to scribe

- Latex stylefile is available on course website
- There is a GoogleDoc where you can sign in
  - Please do so ASAP (so that we have scribes for Wed)
  - Use your Stanford ID to sign in
- Send us the PDF and the LaTeX source by the next class
- Each student is required to scribe at least once
Homeworks

- Homeworks are hard, start early 😊
- Due in the beginning of class (on Mondays)
- 3 late days for the quarter:
  - After late days are used, penalty of 20% per day
- All homeworks must be handed in, even for zero credit
- Late homeworks are handed to submission box in Gates 418
Communication channels

- **Google group:**
  - Main channel for announcements, questions, etc.
  - [http://groups.google.com/group/stanfordcs224w](http://groups.google.com/group/stanfordcs224w)
- **For e-mailing course staff, always use:**
  - cs224w-aut1011-staff@lists.stanford.edu
- **For announcements subscribe to:**
  - cs224w-aut1011-all@lists.stanford.edu
- You can follow us on Twitter: [@cs224w](https://twitter.com/cs224w)
- And use hashtag #cs224w
You are welcome to sit in and audit the class.

Please, send us email saying that you will be auditing the class.

If you’d like to receive announcements, subscribe to the mailing list.
Networks are becoming ubiquitous in science, engineering and beyond

This class should give you the basic foundation for how to think about networks and how to develop new methods and insights

The fun begins...
A map of the Web

What does the Web “look like” at a global level?

Network:
- Nodes = pages
- Edges = hyperlinks

What is a node?
- Problems:
  - Dynamic pages created on the fly
  - “dark matter” – inaccessible database generated
Web as a Graph

I teach a class on Networks.

Networks Course: We have a class blog

Networks Class Blog: This blog post is about Microsoft

Microsoft Home Page
- In early days of the Web links were **navigational**
- Today many links are **transactional**
Other information networks

Citations

References in encyclopedia
How does the Web look like?

- How is the Web linked?
- What is the “map” of the Web?

Web as a directed graph [Broder et al. 2000]:

- A path is a sequence of nodes connected by edges pointing in the right direction
- Given node $v$, what can $v$ reach?
  What can reach $v$?
- Directed graphs – who can reach whom
  - $\text{In}(v) = \{w \mid w \text{ can reach } v\}$
  - $\text{Out}(v) = \{w \mid v \text{ can reach } w\}$
Web as a directed graph

- I'm a student at Univ. of X
- My song lyrics
- Classes
- Networks
- Networks class blog
- Blog post about Company Z
- Univ. of X
- I teach at Univ. of X
- Blog post about college rankings
- I'm applying to college
- USNews College Rankings
- USNews Featured Colleges
Two types of directed graphs:

- **DAG** – directed acyclic graph:
  - Has no cycles: if u can reach v, then v can not reach u

- **Strongly connected**:
  - Any node can reach any node via a directed path

Any directed graph can be expressed in terms of these two types
Strongly connected component (SCC) is a set of nodes $S$ so that:
- Every pair of nodes in $S$ can reach each other
- There is no larger set containing $S$ with this property
Fact: Every directed graph is a DAG on its SCCs.

1. SCCs partition the nodes of G (each node in exactly one SCC)
2. If we build a graph G’ whose nodes are SCCs, and edge between nodes G’ if there is an edge between corresponding SCCs in G, then G’ is a DAG
Example

Example where this representation gives us computational leverage:

- Given a directed graph $G$
- Find the smallest set of nodes $A$ so that every node in $G$ is reachable from at least one node in $A$. 
Graph structure of the Web

- Take a large snapshot of the web and try to understand how it’s SCCs “fit” as a DAG.

- Computational issue:
  - Say want to find SCC containing node \( v \)?
  - Observation:
    - Out\((v)\) ... nodes that can be reached from \( v \)
    - SCC containing \( v \) is:
      \[
      = \text{Out}^{-1}(v) \cap \text{In}^{-1}(v)
      \]
      \[
      = \text{Out}(v; G) \cap \text{Out}(v; G')
      \]
    - where \( G' \) is \( G \) with edge directions reversed.
There is a giant SCC
There won’t be 2 giant SCCs:
  - Just takes 1 page from each to link to one in other – if the components have millions of pages the likelihood of this is large
Broder et al., 2000:
  - Weakly connected component: 90% of the nodes
Bow-tie structure of the Web

- 250 million webpages, 1.5 billion links [Altavista]
What did we learn/not learn about the web?

- **Learn:**
  - Some conceptual organization of the Web (i.e., bowtie)

- **Not learn:**
  - Treats all pages as equal
    - Google’s homepage == my homepage
  - What are the most important pages
    - How many pages have k in-links as a function of k?
      - The degree distribution: $\sim 1/k^{2+\epsilon}$
    - Link analysis ranking -- as done by search engines (PageRank)
  - Internal structure inside giant SCC
    - Clusters, implicit communities? Mini BowTies?
  - How far apart are nodes in the giant SCC:
    - Distance = # of edges in shortest path
    - $\text{Avg} = 16$ [Broder et al.]