Information about the Course

- Lectures: Maria Grineva
- Exercises and Assistance: Kyumars Sheykh Esmaili
- Language: English
- Course webpage: https://www.systems.ethz.ch/education/fs11/struct-social-inf-networks
Lectures and Exercises Schedule

- Lectures: Wed 13:15 - 15:00
- Exercises: Wed 15:15 - 16:00
- No lecture and exercises on 27 Apr (Easter Break)
Exercises

- Kyumars Sheykh Esmaili kyumarss@inf.ethz.ch
- Break up into groups of 2-3 students
- Written exercises tasks every 2 week
- First set of exercises will be given on the next lecture (2nd March); Solutions will be discussed in two weeks (16th March) and so on ...
- Send your solutions to Kyumars during two weeks:
  - Deadline: 12am Wednesday two weeks after exercises were given
Paper Presentations

• List of academic papers related to the course on the course page: https://www.systems.ethz.ch/education/fs11/struct-social-inf-networks/papers

• Every 2 weeks exercise session contains 2 paper presentations (20 mins each)

• Break up into groups of 2-3 students

• Choose a paper and a date for the presentation - see table with dates/presentations on the course page

• Send this information to Kyumars
Exams and Grading

• The exam will be written and in English
• The Exam will be in August (date to be announced)
• Grading:
  60% - written exam
  20% - exercises
  20% - paper presentation
• Exercises and at least one paper presentation are necessary for taking the exam
• David Easley and Jon Kleinberg: “Networks Crowds and Markets. Reasoning about a Highly Connected World”

• You can download the book PDF (find the link on course web page)

• Additional readings - list of papers for presentations at the course webpage: https://www.systems.ethz.ch/education/fs11/struct-social-inf-networks/papers
Should you have any questions

- Kyumars Sheykh Esmaili
  @ CAB F 71.2
  kyumarss@inf.ethz.ch
  - Wednesdays: before or after an exercise session

- Maria Grineva
  @ CAB E 72.3
  maria.grineva@inf.ethz.ch
In this course we will study the structure and behavior of large social and information networks, such as the Web.

Here are the main topics that will be covered in this course:

We begin with the basics of two theories - graph theory and game theory. Graph theory is the study of network structure, while game theory provides models of individual behavior in settings where outcomes depend on the behavior of others. We will then apply both of this theories for network analysis.
In Graph theory part, we will focus on some of the fundamental ideas of social network analysis. The following pictures give hints at some of these ideas.

Pic 1: Corporate email communication network. We can see how the communication is balanced between staying within small groups of the networks and cutting across the groups. This is an example of much more general principle in social networks – strong ties, representing close and frequent social contact, tend to be embedded in tightly-liked regions of the network. While weak ties, representing rare social contact, tend to cross between these regions.

Strong and weak ties principle: So, this principle suggests thinking about a social network as a set of dense groups of strong ties, and these groups interact between each other via weak ties.

Structural holes: Parts of the network that interact very little between each other – we call structural holes.

In this course, we will talk about different advantages for a node (or a person in a social network) to be inside a group or to span a structural hole.

We will also talk about the small-world phenomena. In real-world social networks all people are connected with each other via some surprisingly small number of intermediate steps.
Social network analysis can uncover latent conflicts

The theory of **Structural balance** - reasoning about possible splits of unions inside a network

Latent conflict at Karate club: two non-interacting groups tend to split into two separate parts.

Pic 2: Karate club social network. This example shows how it is possible to find out a latent conflict in a social network. On the pic you can see the social network of friendships within 34-person Karate club. The people labeled 1 and 34 are central, they have many connections with other people, and at the same time they are not friends with each other. And most people are only friends with one or the other of them. These two people were the instructor and the student founder of the club.

You can see the pattern of non-interacting clusters. It is a symptom of a conflict between the two groups of this social network. The symptom predicts that later this network can split into two smaller parts.

Later in this course, we will see how the theory of structural balance can be used to reason about possible splits or unions inside a network.
Course Overview:
Game Theory

- Group of people must simultaneously choose how to act, knowing that outcome will depend on the decisions made by all of them

- Examples:
  - choosing a driving route
  - bidding in an auction (Internet advertising principles)

- Strategy - each player must commit to a strategy

- Payoff - each player receives a payoff depending on the strategies chosen by other players

- Equilibrium - “self-reinforcing” state of a game, when there is no player with an incentive to change their strategy

There are settings in which a group of people must simultaneously choose how to act, knowing that the outcome will depend on the decisions made by all of them.

For example: choosing a driving route during the time when the traffic is heavy,

Another example: is bidding in an auction. If a seller is trying to sell an item via auction, the success of any of the bidder depends not only on how she bids, but also on how everybody else bid. The bidder must take an optimal bidding strategy. In this course, we will talk about bidding auctions for internet advertising. We will see how the system of internet advertising works and its principles.

The goal of this part of the course is to abstract these examples into a common framework. In this framework, a collection of individuals choose a strategy and receive a payoff that depends on the strategies chosen by everyone. So, for our example with choosing a route, strategies for a driver are different routes he can take, and the payoff is the resulting travel time. For bidding auctions: the strategies are the different choices of how to bid, and the payoff is the difference between the price and the real value of the product.

The fundamental part of this framework is the notion of equilibrium. Equilibrium - is a special state of a game when there is no incentive for any player to change her strategy, even if the player knows the strategies of others.
Course Overview:
Markets and Strategic Interaction in Networks

- Graph theory + Game theory = models of behavior in network

Medieval trade routes: the position in the network gives certain economic advantage

Once we have developed graph theory and game theory, we can combine them to produce richer models of behavior on networks. Look for example at this map of medieval trade routes. The network structure encodes a lot about the pattern of trade. The success of participants among other economic factors, depends on the position in the network. The power of a trader depends on the number of connections that he has, and also on the power of the participants he is connected with.
The information on the web has a fundamental network structure.

Links among Web pages, for example, can help us to understand how these pages are related, how they are grouped into different communities, and which pages are the most prominent.

This picture illustrates some of these issues: it shows a network of links among American political blogs. The image and its layout shows the clear separation of the blogging network into two large clusters. These two clusters turn out to correspond to the sets of liberal and conservative blogs respectively.

The network is too large to be presented in a picture. So, the researchers conducted a detailed analysis of the raw linkage data underlying the image. And they found out that it is possible to understand how important and influential the blog within each of these clusters.
Course Overview: Information Networks

- Google - uses network structure to evaluate the quality and relevance of a Web page

- **PageRank**: page is prominent if it receives links from prominent pages. Circularity is a kind of *equilibrium in the link structure*

- Interaction between search engines and authors of Web pages - gives interesting, unexpected effects

- Web is not static, it adapts to new ranking rules: **black SEO (search engine optimization), Content Farms ...**

Current Web search engines such as Google make extensive use of network structure in evaluating the quality and relevance of Web pages. For producing search results, Google evaluates the prominence of a Web page not simply based on the number of links it receives, but based also on its position in the network.

Google’s PageRank idea is that the web page is prominent if it receives links from pages that are themselves prominent; this is a circular kind of notion in which prominence is defined in terms of itself, but we will see that this circularity can be resolved through definitions that are based on a kind of equilibrium in the link structure.

Another interesting example is the interaction between search engines and the authors of Web pages. Whenever Google introduces a new method for evaluating Web pages, deciding which pages to rank highly in its results, the creators of Web content react to this: they optimize what they put on the Web page. So they try to achieve a high rank under the new method.

As a result, during the recent years, Google implicitly has produced a whole market of SEO-based businesses. We can see so called Content Farms - companies that produce huge amounts of low quality content in order to get a high rank on some Google results. A lot of businesses use black SEO to get their sites up on Google: a lot of effort has been put to cheat the algorithm and this of course pollutes the Web.

From this we can understand that changes to a search engine can never be designed under the assumption that the Web will remain static. Rather, the Web will adapt to the ways in which search engines evaluate content. Search methods must be developed with these feedback effects in mind. In this course, we will model and analyze the interaction between search engine and creators of Web pages.
Course Overview:
Network Population Effects

• **Social practices**: ideas, beliefs, opinions, innovations, technologies...

• People **influence** each other’s behavior to spread the practices

• **Information cascades** - individuals adopt an idea to follow a crowd

• **Network effects** or “rich get richer” phenomenon

If we observe a large population over time, we will see patterns by which new ideas, beliefs, opinions, innovations, technologies, products, and social conventions are constantly emerging and evolving. Collectively, we can refer to these as social practices.

The way in which new practices spread through a population depends in large part on the fact that people influence each other’s behavior. In short, as you see more and more people doing something, you generally become more likely to do it as well. We can understand how this happens when look at the network aggregate behavior.

One model of spread of social practices is based on the idea of information cascades. First, you may have some private information about a product on which you base your decision if to use it or not. But if you see many people use it, it is natural to assume that they too have their own information.

In case of Web sites like YouTube, the fact that many people use it can suggest that these people know something about its quality, and new people start using it because of that reason. Similarly, seeing that a restaurant is extremely crowded every weekend, can suggest that many people think it is a good place.

In the result, new people often do not rely on private information, and only rely on the fact that many people use the product. And this produces so called information cascade. Even a rational individual chooses not to take into account his private information, but follow a crowd.

Another model is so called network effect. For example with YouTube, regardless of whether YouTube has better or worse features compared to its competitors, once it has become the most popular video sharing site, there is an added value in using it. Such network effects make the success of product and technologies that are already doing well - even more stable; in a market where network effects are at work, the leader can be hard to displace.
Course Overview: Network Structural Effects

- Taking into account the **network structure** to understand how people influence each other
- **Cascading behavior** in networks

As we have just seen, the question of how people influence each other’s behavior is already interesting even when the actual structure of the underlying network is left implicit. But taking network structure into account can give us even more insights into how such kinds of influence take place.

When individuals have incentives to adopt the behavior of their neighbors in the network, we can get cascading effects, where a new behavior starts with a small set of initial adopters, and then spreads outward through the network.

Look for example, at this picture on the left we can see how email recommendations of a new book spread from the four initial readers.

The right picture shows the spread of an epidemic disease. There are fundamental differences in processes of how a new node adopts the recommendations or becomes affected. But at the level of the network, the dynamics of spread are similar.
So now let us proceed to the first part of the course.

I will provide some of the basic ideas behind graph theory. This will allow us to formulate basic network properties in a unifying language. The central definitions here are simple enough that we can describe them relatively quickly; after this, we consider some fundamental applications of these definitions.
A graph is a way of specifying relationships among a collection of items. A graph consists of a set of objects, called nodes, with certain pairs of these objects connected by links called edges. For example, the graph in the picture consists of 4 nodes labeled A, B, C, and D, with B connected to each of the other three nodes by edges, and C and D connected by an edge as well. We say that two nodes are neighbors if they are connected by an edge.

In the picture on the left, the relationship between the two ends of an edge are symmetric; the edge simply connects them to each other. In many settings, however, we want to express asymmetric relationships, for example, that A points to B. For this purpose, we define a directed graph to consist of a set of nodes, as before, together with a set of directed edges; each directed edge is a link from one node to another, with the direction being important.
Graphs as Models of Networks

- Graphs are useful because they serve as mathematical models of network structures.

The picture shows the network structure of the Internet - then called the Arpanet - in December 1970.
Paths and Connectivity

- **Path** is a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge.

- **Cycle** is a path with at least three edges, in which the first and last nodes are the same, but otherwise all nodes are distinct.

- A graph is *connected* if, for every pair of nodes, there is a path between them.

We now turn to some of the fundamental concepts and definitions surrounding graphs.

A path is simply a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge. Sometimes it is also useful to think of the path as containing not just the nodes but also the sequence of edges linking these nodes.

More precisely, a cycle is a path with at least three edges, in which the first and last nodes are the same, but otherwise all nodes are distinct.

Given a graph, it is natural to ask whether every node can reach every other node by a path. With this in mind, we say that a graph is connected if for every pair of nodes, there is a path between them.
Paths and Connectivity

• If a graph is not connected, it breaks apart into *components*

• Large real-world networks often have a *giant component* - a connected component that contains a significant fraction of all nodes

Giant Component: Consider the social network of the entire world, with a link between two people if they are friends. First, is this global friendship network connected? Presumably not. After all, connectivity is a fairly brittle property, in that the behavior of a single node (or a small set of nodes) can negate it. Even though the global friendship network may not be connected, the component you inhabit seems very large indeed: it reaches into most parts of the world, includes people from many different backgrounds, and seems in fact likely to contain a significant fraction of the world’s population.

This is true when one looks across a range of network datasets: large, complex networks often have what is called a giant component. Giant component is an informal term for a connected component that contains a significant fraction of all the nodes. Moreover, when a network contains a giant component, it almost always contains only one.
A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted.

An interesting example is depicted in Figure 2.7, which shows the romantic relationships in an American high school over an 18-month period. You can see a large giant component. Many students who had only one relationship, are included into this big company.
Distance and Breadth-First Search

- **Length of a path** - the number of steps it contains from beginning to end
- **Distance** between two nodes - the length of the shortest path between them
- How to figure out a distance between two nodes: *breadth-first search*

In addition to simply asking whether two nodes are connected by a path, it is also interesting in most settings to ask how long such a path is in transportation, Internet communication, or the spread of news and diseases, it is often important whether something owing through a network has to travel just a few hops or many.

We define the distance between two nodes in a graph to be the length of the shortest path between them.
Breadth-first Search

- searches the graph outward from a starting point

The most natural way to do this formalized by the technique called breadth-first search.

(1) You first declare all of your actual friends to be at distance 1.
(2) You then find all of their friends (not counting people who are already friends of yours), and declare these to be at distance 2.
(3) Then you find all of their friends (again, not counting people who you’ve already found at distances 1 and 2) and declare these to be at distance 3.

(….) Continuing in this way, you search in successive layers, each representing the next distance out. Each new layer is built from all those nodes that (i) have not already been discovered in earlier layers, and that (ii) have an edge to some node in the previous layer.

This technique is called breadth-first search, since it searches the graph outward from a starting node, reaching the closest nodes first. In addition to providing a method of determining distances, it can also serve as a useful conceptual framework to organize the structure of a graph, arranging the nodes based on their distances from a fixed starting point.
Small-world Phenomenon or Six Degrees of Separation

- **Small-world phenomenon** - the idea that the world looks “small” when you think of how short a path of friends it takes to get from you to almost anyone else

- Microsoft Instant Messenger network: average distance = 6.6

If we go back to our thought experiments on the global friendship network, we see that the argument explaining why you belong to a giant component in fact asserts something stronger: not only do you have paths of friends connecting you to a large fraction of the world’s population, but these paths are surprisingly short.

This idea is called the small-world phenomenon - the idea that the world looks “small” when you think of how short a path of friends it takes to get from you to almost anyone else. It’s also known as the six degrees of separation;

Interesting experiment on understanding this phenomena has been conducted using Microsoft Instant Messaging data. Researchers analyzed the 240 million active user accounts on Microsoft Instant Messenger, building a graph in which each node corresponds to a user, and there is an edge between two users if they engaged in a two-way conversation at any point during a month-long observation period.

This graph turned out to have a giant component containing almost all of the nodes, and the distances within this giant component were very small. An estimated average distance of 6.6.

The picture shows the distribution of distances averaged over a random sample of 1000 users: breadth-first search was performed separately from each of these 1000 users, and the results from these 1000 nodes were combined to produce the plot in the figure. The reason for this estimation by sampling users is a computational one: the graph was so large that performing breadth-first search from every single node would have taken an astronomical amount of time.
Small-world Phenomenon or Six Degrees of Separation

- Collaboration between mathematicians.
  Central - Paul Erdos

- Most mathematicians have Erdos-number of at most 4-5

Researchers have also discovered very short paths in the collaboration networks within professional communities. In the domain of mathematics, for example, there is a well know mathematician Paul Erdos I who published around 1000 papers over his career. Paul Erod was taken as a central point in the collaborative structure of the field.

In the picture, nodes correspond to mathematicians, and edges connecting pairs who have jointly authored a paper. Pic shows a small hand-drawn piece of the collaboration graph, with paths leading to Paul Erdos.

Now, a mathematician's Erdos number is the distance from him or her to Erdos in this graph. The point is that most mathematicians have Erdos numbers of at most 4 or 5. For example, Albert Einstein's has erdos number 2, Enrico Fermi's is 3,
Another well known experiment is a collaboration graph of movie actors and actresses: nodes are movie actors, an edge connects two actors if they've appeared together in a movie. An actor's Bacon number is his or her distance in this graph to Kevin Bacon.

Using cast lists from the Internet Movie Database (IMDB), it is possible to compute Bacon numbers for all performers via breadth-first search and as with mathematics, it's a small world indeed. The average Bacon number, over all performers in the IMDB, is approximately 2.9, and it's a challenge to find one that's larger than 5.
Triadic Closure

• Most networks are not static

• Useful to think about *how a network evolves over time*: how nodes arrive and depart, edges form and vanish
Triadic Closure

- If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends two

- **Triadic closure** - a principle

- The formation of the edge between B and C illustrates triadic closure

- B and C have an *incentive* to become friends (form a BC edge)

Triadic closure is a basic principle that explains forming new edges in a network.

Speaking in terms of social network of friend relationships, we can say: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends two

This principle is illustrated in the picture if nodes B and C have a friend A in common, then the formation of an edge between B and C produces a situation in which all three nodes A, B, and C have edges connecting each other - a structure we refer to as a triangle in the network. The term “triadic closure” comes from the fact that the B-C edge has the effect of “closing” the third side of this triangle.
The Clustering Coefficient

- **The clustering coefficient** of a node A is a probability that two randomly selected friends of A are friends with each other.

- Example:
  the clustering coeff of A before new edges form = \(\frac{1}{2}\), after = \(\frac{1}{6}\)

One of these is the clustering coefficient. The clustering coefficient of a node A is defined as the probability that two randomly selected friends of A are friends with each other. In other words, it is the fraction of pairs of A’s friends that are connected to each other by edges. For example, the clustering coefficient of node A in the picture on the left is \(\frac{1}{6}\) (because there is only the single C-D edge among the six pairs of friends B-C, B-D, B-E, C-D, C-E, and D-E, and it has increased to \(\frac{1}{2}\) in the second picture because there are now the three edges B-C, C-D, and D-E among the same six pairs. In general, the clustering coefficient of a node ranges from 0 (when none of the node’s friends are friends with each other) to 1 (when all of the node’s friends are friends with each other), and the more strongly triadic closure is operating in the neighborhood of the node, the higher the clustering coefficient will tend to be.
The Strength of Weak Ties

- An edge that joins two nodes A and B in a graph is called a **bridge** if deleting the edge would cause A and B to lie in two different components.
- **Local bridge** - in real-world networks (with a giant component) - if deleting an edge between A and B would increase distance > 2.

Here is another pair of useful definitions. Look at the left picture.

The person labeled A has four friends in this picture, but one of her friendships is qualitatively different from the others: A’s links to C, D, and E connect her to a tightly-knit group of friends who all know each other, while the link to B seems to reach into a different part of the network.

To make precise the sense in which the A-B link is unusual, we introduce the following definition. We say that an edge joining two nodes A and B in a graph is a bridge if deleting the edge would cause A and B to lie in two different components. In other words, this edge is literally the only route between its endpoints, the nodes A and B.

Now, let’s remember our discussion about giant components and small-world properties. Bridges are extremely rare in real social networks. You may have a friend from a very different background, and it may seem that your friendship is the only thing that bridges your world and his, but one expects in reality that there will be other, hard-to-discover, multi-step paths that also span these worlds. In other words, if we were to look at the left picture as it is embedded in a larger, real social network, we would likely see a picture that looks like the one on the right. Here, the A-B edge isn’t the only path that connects its two endpoints; though they may not realize it, A and B are also connected by a longer path through F, G, and H. We say that an edge joining two nodes A and B in a graph is a local bridge if its endpoints A and B have no friends in common. In other words, if deleting the edge would increase the distance between A and B to a value strictly more than two.

Local bridges, especially those with reasonably large span, still play roughly the same role that bridges do, though in a less extreme way — they provide their endpoints with access to parts of the network, and hence sources of information, that they would otherwise be far away from.
The Strong Triadic Closure Property

- Different level of strength of links in social network. For simplicity: **strong** and **weak**
- A node $A$ violates the Strong Triadic Closure property if it has strong ties to two other nodes $B$ and $C$ and there is no edge at all between $B$ and $C$.
- $A$ satisfies the STC if it does not violate it

**If $A-F$ were a strong tie, then $A$ and $F$ would both violate STC**

we must be able to distinguish between different levels of strength in the links of a social network.

We are not going to define "strength" precisely, but we just categorize all links in the social network as belonging to one of two types: strong ties (the stronger links, corresponding to friends), and weak ties (the weaker links, corresponding to acquaintances)

Once we have decided on a classification of links into strong and weak ties, we can take a social network and annotate each edge with a designation of it as either strong or weak. For example, look at this annotated network on the slide.

It is useful to go back and think about triadic closure in terms of this division of edges. This suggests the following qualitative assumption:

If a node $A$ has edges to nodes $B$ and $C$, then the $B-C$ edge is especially likely to form if $A$'s edges to $B$ and $C$ are both strong ties.

We say that a node $A$ violates the Strong Triadic Closure Property if it has strong ties to two other nodes $B$ and $C$, and there is no edge at all (either a strong or weak tie) between $B$ and $C$. We say that a node $A$ satisfies the Strong Triadic Closure Property if it does not violate.

You can check that no node in the picture violates the Strong Triadic Closure Property, and hence all nodes satisfy the Property. On the other hand, if the $A-F$ edge were to be a strong tie rather than a weak tie, then nodes $A$ and $F$ would both violate the Strong Triadic Closure Property: Node $A$ would now have strong ties to nodes $E$ and $F$ without there being an $E-F$ edge, and node $F$ would have strong ties to both $A$ and $G$ without there being an $A-G$ edge.
Local Bridges and Weak Ties

- Claim: if a node $A$ in a network satisfies the Strong Triadic Closure property and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.

- The argument proceeds by contradiction.

We're going to justify this claim as a mathematical statement.

The argument is actually very short, and it proceeds by contradiction. Take some network, and consider a node $A$ that satisfies the Strong Triadic Closure Property and is involved in at least two strong ties. Now suppose $A$ is involved in a local bridge I say, to a node $B$ that is a strong tie. We want to argue that this is impossible. First, since $A$ is involved in at least two strong ties, and the edge to $B$ is only one of them, it must have a strong tie to some other node, which we'll call $C$. Now let's ask: is there an edge connecting $B$ and $C$? Since the edge from $A$ to $B$ is a local bridge, $A$ and $B$ must have no friends in common, and so the $B$-$C$ edge must not exist. But this contradicts Strong Triadic Closure.
Tie Strength in Large-Scale Data

• Experimental study on “who-talks-to-whom” network data maintained by a cell phone provider

• edge strength = total number of minutes of talk

• soften the definition of local bridges:

\[
\frac{\text{number of nodes who are neighbors of both } A \text{ and } B}{\text{number of nodes who are neighbors of at least one of } A \text{ or } B'}
\]
Using these definitions, we can ask how the neighborhood overlap of an edge depends on its strength; the strength of weak ties predicts that neighborhood overlap should grow as tie strength grows. In fact, this is borne out extremely cleanly by the data.

Figure 3.7 shows the neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. Thus, as we go to the right on the \( x \)-axis, we get edges of greater and greater strength, and because the curve rises in a strikingly linear fashion, we also get edges of greater and greater neighborhood overlap. The relationship between these quantities thus aligns well with the theoretical prediction.
It is interesting to study tie strength on social network sites.

When we see people maintaining hundreds of friendship links on a social-networking site, we can ask how many of these correspond to strong ties that involve frequent contact, and how many of these correspond to weak ties that are activated relatively rarely.

The study was performed on friendship network of Facebook. The picture shows the network neighborhood of a sample Facebook user, consisting of all his friends, and all links among his friends. The picture in the upper-left shows the set of all declared friendships in this user's profile; the other three pictures show how the set of links becomes sparser once we consider only maintained relationships, one-way communication, or reciprocal communication.

Moreover, as we restrict to stronger ties, certain parts of the network neighborhood thin out much faster than others. For example, in the neighborhood of the sample user in the picture, we see two distinct regions where there has been a particularly large amount of triadic closure: one in the upper part of the drawing, and one on the right-hand side of the drawing. However, when we restrict to links representing communication or a maintained relationship, we see that a lot of the links in the upper region survive, while many fewer of the links in the right-hand region do. One could conclude that the right-hand region represents a set of friends from some earlier phase of the user's life (perhaps from school) who declare each other as friends, but do not actively remain in contact; the upper region, on the other hand, consists of more recent friends (perhaps co-workers) for whom there is more frequent contact.
We can make the relative abundance of these different types of links quantitative through the plot in Figure 3.9. On the x-axis is the total number of friends a user declares, and the curves then show the (smaller) numbers of other link types as a function of this total. There are several interesting conclusions to be drawn from this. First, it confirms that even for users who report very large numbers of friends on their profile pages (on the order of 500), the number with whom they actually communicate is generally between 10 and 20, and the number they follow even passively (e.g. by reading about them) is under 50.
Closure and Structural Holes

- in social networks, some nodes are positioned at the interfaces between multiple groups, while others in the middle of a single group

- **A** - has high clustering coefficient

- **Embeddedness** of an edge - a number of common neighbors shared by the two endpoints

- **A-B** has embeddedness=2

Now what if we ask about different roles a node can play depending on his position in a social network.

In social networks, access to edges that span different groups is not equally distributed across all nodes: some nodes are positioned at the interface between multiple groups, with access to boundary-spanning edges, while others are positioned in the middle of a single group.

There is different experiences that nodes have in a network like the one in the picture. Particularly, if look at nodes A and B. A sits at the center of a single tightly-knit group, and node B sits at the interface between several groups.

**Embeddedness.** Let's start with node A. Node A's set of network neighbors has considerable triadic closure; A has a high clustering coefficient. That is, the fraction of pairs of neighbors who are themselves neighbors is high for A. To talk about the structure around A it is useful to introduce an additional definition.

We define the embeddedness of an edge in a network to be the number of common neighbors the two endpoints have. Thus, for example, the A-B edge has an embeddedness of two, since A and B have the two common neighbors E and F. Now we can say that local bridges are precisely the edges that have an embeddedness of zero, since they were defined as those edges whose endpoints have no neighbors in common.

In the example shown in picture, what noticable about A is the way in which all of his edges have significant embeddedness. A long line of research in sociology has argued that if two individuals are connected by an embedded edge, then this makes it easier for them to trust one another. Indeed, the presence of mutual friends puts the interactions between two people "on display" in a social sense, even when they are carried out in private; in the event of misbehavior by one of the two parties to the interaction, there is the potential for social sanctions and reputational consequences from their mutual friends.

This is absolutely different for B. B has edges with zero ebmbeddedness. In this respect, the interactions that B has with C and D are much riskier than the embedded interactions that A experiences, since there is no one who knows both people involved in the interaction.
Closure and Structural Holes

- **B** spans a `structural hole` - the `empty space` in the network between two sets of nodes that do not otherwise interact closely
- Studied by Ron Burt

- **B** has early access to information that originates in multiple noninteracting parts of the network
- **B** has an opportunity for novel ideas by combining the different sources of information
- **B** is a social `gatekeeper`. She regulates the access of C and D to her group

Structural holes. Thus far we have been discussing the advantages for a node A in the picture from the closure in his network neighborhood. But a related line of research in sociology has shown that network positions such as node B’s, at the ends of multiple local bridges, has a distinct set of equally fundamental advantages.

The canonical setting for this argument is the social network within an organization or company, consisting of people who are in some ways collaborating on common objectives and in other ways implicitly competing for career advancement. Empirical studies of managers in large corporations has correlated an individual’s success within a company to their access to local bridges.

B, with her multiple local bridges, spans a structural hole in the organization - the “empty space” in the network between two sets of nodes that do not otherwise interact closely. The argument is that B’s position offers advantages in several dimensions relative to A’s.

The first advantage is an informational one: B has early access to information originating in multiple, non-interacting parts of the network.

A second, related kind of advantage is that standing at one end of a local bridge can boost creativity of that person. Experience from many domains suggests that innovations often arise from the unexpected synthesis of multiple ideas, each of them on their own perhaps well-known, but well-known in distinct and unrelated fields. Thus, B’s position at the interface between three non-interacting groups gives her not only access to the combined information from these groups, but also the opportunity for novel ideas by combining these disparate sources of information in new ways.

Finally, B’s position in the network provides an opportunity for a kind of social “gatekeeping” she regulates the access of both C and D to the tightly-knit group she belongs to, and she controls the ways in which her own group learns about information coming from C’s and D’s groups.
As we have seen, real world social networks often have “tightly-knit regions” with strong ties inside them, connected via weak ties. Now the question is how we can automatically identify these tightly-knit communities inside a network.

This will be our focus here: describing a method that can take a network and break it down into a set of tightly-knit regions, with sparser interconnections between the regions. We will refer to this as the problem of graph partitioning, and the constituent parts the network is broken into as the regions arising from the partitioning method.

To give a sense for what we might hope to achieve from such a method, let’s consider the example. On the slide you can see a social network of a karate club that we already discussed before in this lecture. A conflict among the club members led the club to split into two clubs. The picture shows the network structure, with the membership in the two clubs after the division indicated by the shaded and unshaded nodes. The two conflicting groups here are still heavily interconnected, so we need an algorithm that would measure the strength of the edges that connect the two groups.
Many different approaches have been developed for the problem of graph partitioning and they can be divided into two main groups.

One class of methods focuses on identifying and removing the "spanning links" between densely-connected regions. Once these links are removed, the network begins to fall apart into large pieces; within these pieces, further spanning links can be identified, and so on this process continues. We will refer to these as divisive methods of graph partitioning, since they divide the network up as they go.

An alternate class of methods starts from the opposite end of the problem. We can say it is a bottom up approach. Such methods find nodes that are likely to belong to the same region and merge them together. Once this is done, the network consists of a large number of merged chunks; the process then looks for chunks that should be further merged together. We refer to these as agglomerative methods of graph partitioning, since they glue nodes together into regions as they go.
To illustrate the conceptual differences between these two kinds of approaches, let's consider the simple graph in this picture. You can see a broad separation between one region consisting of nodes 1-7, and another consisting of nodes 8-14. Within each of these regions, there is a further split: on the left into nodes 1-3 and nodes 4-6; on the right into nodes 9-11 and nodes 12-14. Note how this example already illustrates that the process of graph partitioning can be viewed as nested steps: larger regions potentially contain several smaller, even more tightly-knit regions “nested” within them.

In fact, a number of graph partitioning methods will find the nested set of regions indicated in picture. Divisive methods will generally proceed by breaking apart the graph first at the 7-8 edge, and subsequently at the remaining edges into nodes 7 and 8. Agglomerative methods will arrive at the same result from the opposite direction, first merging the four triangles into clumps, and then finding that the triangles themselves can be naturally paired off.
A Method for Graph Partitioning: The Notion of Betweenness

- First removing edges with heavy traffic: bridges and local bridges
- **Betweenness** - a measure of a “traffic” of an edge

A first idea, that comes to mind, is that since bridges and local bridges often connect weakly interacting parts of the network, we should try removing these bridges and local bridges first.

We define the betweenness of an edge to be the total amount of flow it carries, counting flow between all pairs of nodes using this edge.
For example, we can determine the betweenness of each edge in the picture as follows.

Let’s first consider the 7-8 edge. For each node A in the left half of the graph, and each node B in the right half of the graph, their full unit of flow passes through the 7-8 edge. On the other hand, no flow passing between pairs of nodes that both lie in the same half uses this edge. As a result, the betweenness of the 7-8 edge is $7 	imes 7 = 49$.

The 3-7 edge carries the full unit of flow from each node among 1, 2, and 3 to each node among 4-14. Thus, the betweenness of this edge is $3 \times 11 = 33$. The same goes for the edges 6-7, 8-9, and 8-12.

The 1-3 edge carries all the flow from 1 to every other node except 2. As a result, its betweenness is 12. Similarly, the other edges linked from 3, 6, 9, and 12 into their respective triangles have betweenness 12 as well.

Finally, the 1-2 edge only carries flow between its endpoints, so its betweenness is 1. This also holds for the edges 4-5, 10-11, and 13-14.
Grivan-Newman Method for Graph Partitioning

- The Girvan-Newman method: Successively delete edges of high betweenness

Edges of high betweenness are the ones that carry the highest volume of traffic along shortest paths. It is natural to try removing these first.

This is the main idea of the Girvan-Newman method, which can now be summarized as follows.

1. Find the edge of highest betweenness, or multiple edges of highest betweenness, and remove these edges from the graph. This may cause the graph to separate into multiple components. If so, this is the first level of regions in the partitioning of the graph.
2. Now recalculate all betweennesses, and again remove the edges of highest betweenness. This may break some of the existing components into smaller components; if so, these are regions nested within the larger regions.

(…) Proceed in this way as long as edges remain in graph, in each step recalculating all betweennesses and removing the edge or edges of highest betweenness.

Thus, as the graph falls apart first into large pieces and then into smaller ones, the method naturally exposes a nested structure in the tightly-knit regions.

The sequence of steps in Figure 3.17 in fact exposes some interesting points about how the method works.

When we calculate the betweennesses in the first step, the 5-7 edge carries all the flow from nodes 1-5 to nodes 7-11, for a betweenness of 25. The 5-6 edge, on the other hand, only carries flow from node 6 to each of nodes 1-5, for a betweenness of 5. (Similarly for the 6-7 edge.)

Once the 5-7 edge is deleted, however, we recalculate all the betweennesses for the second step. At this point, all 25 units of flow that used to be on this deleted edge have shifted onto the path through nodes 5, 6, and 7, and so the betweenness of the 5-6 edge (and also the 6-7 edge) has increased to $5 + 25 = 30$. This is why these two edges are deleted next.
Networks in Their Surrounding Contexts

- A context in which a social network is embedded has effects on its structure

- **surrounding contexts**: factors that exist outside the nodes and edges of a network, but which affect how the network's structure evolves

So far, we focused primarily on the network as an object of study in itself, relatively independent of the world in which it exists. However, the contexts in which a social network is embedded will generally have significant effects on its structure. Each individual in a social network has a distinctive set of personal characteristics, and similarities two people's characteristics can strongly influence whether a link forms between them.

So, now let us consider how network's surrounding contexts: factors that exist outside the nodes and edges of a network, can affect how the network's structure evolves.
Homophily

- **Homophily** - the principle that we tend to be similar to our friends
- **Viewed collectively**, your friends are generally similar to you

Social network from a town’s middle school and high school.
Groups form base on race and on the school.

One of the most basic notions in the structure of social networks is homophily - the principle that we tend to be similar to our friends. Typically, your friends don't look like a random sample of the underlying population: viewed collectively, your friends are generally similar to you.

Often, when we look at a network, homophily is a dominant feature of its overall structure. For example, the picture shows the social network within a town's middle school and high school. In this picture, students of different races are drawn as differently-colored circles. Two dominant divisions within the network are apparent. One division is based on race; the other, based on age and school, separates students in the middle school from those in the high school.

There are many other structural details in this network, but the effect homophily stands out when the network is viewed at a global level.
Measuring Homophily

- Is this network exhibit **homophily**?
- Suppose:
  - fraction $p$ - male, $q$ - female;
  - randomly assign each node a gender with probabilities $p$ and $q$;
  - $p^2$ - probability both ends of an edge are male, $q^2$ - probability both ends are female
  - $2pq$ - cross gender edge (one end - male, other female)
- **Homophily test**: if the fraction of cross-gender edges is significantly less than $2pq$, then there is evidence for homophily

- 5 of 18 edges are cross-gender
- $p=2/3$, $q=1/3$, $2pq=8/18$
- $5/18 < 8/18 \Rightarrow$ yes, there is homophily

---

Thursday, February 24, 2011

How to understand if homophily is present in the network and how to measure it?
To make this question concrete, we need to formulate it more precisely: given a particular characteristic of interest (like race, or age), is there a simple test we can apply to a network in order to estimate whether it exhibits homophily according to this characteristic?

Since the example on the last slide is too large to inspect by hand, let's consider this question on a smaller example where we can develop some intuition. Let's suppose in particular that we have the friendship network of an elementary-school classroom, and we suspect that it exhibits homophily by gender: boys tend to be friends with boys, and girls tend to be friends with girls. For example, the graph in the picture shows the friendship network of a (small) hypothetical classroom in which the three shaded nodes are girls and the six unshaded nodes are boys.

There is a natural numerical measure of homophily that we can use to address questions.

Suppose we have a network in which a $p$ fraction of all individuals are male, and a $q$ fraction of all individuals are female. Consider a given edge in this network. If we independently assign each node the gender male with probability $p$ and the gender female with probability $q$, then both ends of the edge will be male with probability $p^2$, and both ends will be female with probability $q^2$. On the other hand, if the first end of the edge is male and the second end is female, or vice versa, then we have a cross-gender edge, so this happens with probability $2pq$.

So we can summarize the test for homophily according to gender as follows: If the fraction of cross-gender edges is significantly less than $2pq$, then there is evidence for homophily.

In the picture, for example, 5 of the 18 edges in the graph are cross-gender. Since $p = 2/3$ and $q = 1/3$ in this example, we should be comparing the fraction of cross-gender edges to the quantity $2pq$. In other words, with no homophily, one should expect to see 8 cross-gender edges rather than 5, and so this example shows some evidence of homophily.

There are a few points to note here. First, the number of cross-gender edges in a random assignment of genders will deviate some amount from its expected value of $2pq$, and so to perform the test in practice one needs a working definition of a threshold when you decide there is a homophily.

It's also easily possible for a network to have a fraction of cross-gender edges that is significantly more than $2pq$. In such a case, we say that the network exhibits inverse homophily. Recall the network of romantic relationships for example would exhibit inverse homophily.
• This lecture:
  Chapters 1; 2.1-2.3; 3.1-3.6; 4.1 of the text book

• Next lecture:

  Structural balance in Networks;
  Introduction to Game Theory
Advertisement:

Accenture Internships in India

- If you are interested in an internship in India, please, contact: donaldk@inf.ethz.ch
- Duration: June - September 2011 (no classes)
- Places: Bangalore, Chennai oder Hyderabad
- Requirements:
  - doing well with regard to your study plan
  - English
  - Programming languages: Java, SQL
  - Bachelor and Master students
- Application deadline: **15. March 2011 (by E-Mail)**
- Donald Kossmann:
  "This is an adventure!!! I would do it myself, if I met the requirements."
- „Session Exams“ can be taken in India