Large-Scale Networks

2 – Graph Ties
Networks and graphs

- A network often denotes a real system
  - With nodes connected by links

- A graph is a mathematical representation
  - With vertices connected by edges

- These terms are often used interchangeably

- A graph $G(N, E)$ has a set of vertices $N$ and a set of edges $E$. 
Mathematical representation

- Arpanet (former Internet) in 1970

- A simplified graph representation of Arpanet
Graph examples

Actor 1
- Movie 1
- Movie 2

Actor 2
- Movie 3

Actor 3

Actor 4

Peter
- friend

Mary
- co-worker

Tom

Albert
- brothers

Protein 1

Protein 2

Protein 5

Protein 9

| N | = 4
| E | = 4
**Undirected vs. directed graphs**

- **Undirected graph**
  - Ex: Facebook friendship

- **Directed graph**
  - Ex: The world wide web
Paths and connectivity

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

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

- **Connectivity**: a graph is *connected* if for every pair of nodes, there is a path between them
Components

- A **connected component** of a graph (often shortened just to the term “component”) is a subset of the nodes such that:
  - every node in the subset has a path to every other; and
  - the subset is not part of some larger set with the property that every node can reach every other.

The graph on the right consists of three such pieces: one consisting of nodes A and B, one consisting of nodes C, D, and E, and one consisting of the rest of the nodes.
Components

- The collaboration graph of the biological research center Structural Genomics of Pathogenic Protozoa (SGPP).

- How many connected components does it have?
Components

- The collaboration graph of the biological research center Structural Genomics of Pathogenic Protozoa (SGPP).

- How many connected components does it have?

- It consists of three distinct connected components
Giant component

Qualitative thinking about connected components of large scale networks

- Consider the social network world-wide

- Is it connected? Presumably not
  - A person with no living friends will be a one-node component

- You have probably friends in a foreign country
  - You are then in the same component as them
  - Their friends and descendants are also in the same component

- Intuitively, even though the network is not connected, it probably has a giant component: a connected component with a large fraction of all nodes
Giant component

When a network contains a giant component it contains generally only one

- Try to imagine that there were two giant components, each with hundreds of millions of people

- All it would take is a single edge from someone in the first of these components to someone in the second to become one giant component

- It is generally unconceivable that such edge would not form!
Merge of giant components

The merge of giant components may have dramatic consequences

- Think about the colonisation of Australia (1788)
  - One could see British and Australians as two giant connected components that co-existed for a long time
  - Human diseases evolved independently
  - Upon colonization the two components merged

- Series of European diseases such as
  - Measles
  - Smallpox
  - Tuberculosis

- A smallpox epidemic in 1789 is estimated to have killed up to 90% of the Darug people
Components

- 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.
Components

- 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.

- What do you observe?
Components

- 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.

- What do you observe?
Components

- The large component plays an important role in the spread of sexually transmitted diseases.

- Take a student who had a single partner during the study.

- Without knowing it, he is likely part of the large component.

⇒ So, he may be part of many paths of potential transmissions.
Distance
Distance

- Besides connectivity through a path, the length of the path is important.

- The length of a path is the number of steps it contains from beginning to end.
  = number of edges the path contains.

- The distance between two nodes in the graph is the length of the shortest path between them.
  - If the distance from $u$ to $v$ is 3 then there is no path between them with length < 3.

- The diameter of the graph is the longest distance among all pairs of nodes.

- Determining the distance may seem easy on small networks, but requires a systematic way for large-scale networks.
Distance

**Breadth-First Search**: searching network from starting node, closest nodes first

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.
Distance

Breadth-First Search discovers layers one at a time from the starting node downward.

- **distance 1**
  - your friends
- **distance 2**
  - friends of friends
- **distance 3**
  - friends of friends of friends
- all nodes, not already discovered, that have an edge to some node in the previous layer
Distance

Result of Breadth-first-search starting at MIT on the 1970 ARPANET network
The Small World Phenomenon
The small-world phenomenon

- The reasons why you belong to a giant component of a social network
  - There are path of friends connecting you to a large fraction of the population
  - These paths are surprisingly short: people you don’t know are length 2 apart
- The *small-world phenomenon* is the idea that the world “looks” small

- Also known as the “six degree of separation”

- In the 1960’s Stanley Milgram did a $680 experiment to test the idea that people are really connected in the global friendship by short chains of friends, i.e., the distance between any pair of people is remarkably short.
The small-world phenomenon

Stanley Milgram conducted the first significant empirical study of the small world phenomenon [Mil67]

- He asked randomly chosen “starter” individuals to try forwarding a letter to a designated “target” person living in Sharon, MA, USA.

- He provided:
  - The target name
  - Address
  - Occupation
  - Personal information

- The participants could not mail the letter directly to the target

- They could only forward the letter to an acquaintance that (s)he knew on a first name basis with the goal of reaching the target as fast as possible
The small-world phenomenon

- 1/3 of the letters eventually arrived at the target
- The median of the number of steps was 6
- This served as an experimental evidence for the existence of short paths
- This style of experiments were reproduced by other people

- The experiment showed:
  1. There are many short paths
  2. People are effective at collaboratively finding these short paths
The small-world phenomenon

Distribution of path lengths among the sixty-four chains that succeeded in reaching the target. The median length was 6.
The small-world phenomenon

- The fact that so many letters reached their destination by so short paths was striking

- There are few caveats though:
  - The six degree of separation does not apply to all pairs of nodes in the network
  - The paths were to a single and fairly affluent (known) target
  - Many letters never arrived
  - Does short distance mean that socially close?
The small-world phenomenon

- The reason why people consider social networks as small worlds is that:
  - This notion was proposed by Milgram but also
  - This was confirmed in settings where we have the full network structure

- To get the distances of these given network structures
  - Code the Breadth-First Search algorithm and load the structure in a computer
  - Run the algorithm on the structure and extract the distances
The small-world phenomenon

- One of the largest computational study analyzed 240M active Microsoft Instant Messaging users [LH08]
  - A node represents a user
  - There is an edge between u and v if they engaged in a two-way conversation during a month-long observation period

- Conclusions:
  - The network contains a giant component containing almost all the nodes
  - The distances within this component were very small
    - Average distance of 6.6 and median distance of 7

- Remarks:
  - The experiment is quite different from Milgram’s
  - Users are technologically endowed and conversation does not imply friendship
The small-world phenomenon

Distribution of distances averaged over a random sample of 1000 users of the graph of all active Microsoft Instant Messenger user accounts during one month.
Conclusion

- A network is a graph with nodes and edges

- Large-scale networks are hard to analyze
  - Manually is impossible
  - Sometimes too large for computers as well

- Properties of networks are important for analysis
  - Path, length, distance
  - Connectivity, connected component

- Intuition tells us that large social networks tends to have:
  - One giant connected component (if not yet, it is likely that merging will occur)
  - Other connected components tend to be much smaller
  - The median/average distances between nodes is often surprisingly small
Triadic Closure
On the power of acquaintances

- In the 70's Mark Granovetter interviewed people who had recently changed employers to learn how these people had discovered their new jobs [Gra74]

- Personal contact led many people to their current job

- More acquaintances rather than close friends

- While your closest friends have more motivation to help you, why are your acquaintances so much more helpful?
On the power of acquaintances

- **Structural conclusion:**
  - These acquaintances span different regions of the network

- **Interpersonal conclusion:**
  - There are consequences in the friendship of two people depending on their strength
Clustering coefficient

- The *triadic closure* principle:
  - If two people in a social network have a friend in common, then there is an increased probability that they will become friends at some point in the future [Rap53]

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

- The clustering coefficient is a social network metric to measure the prevalence of triadic closure.
Clustering coefficient

If node 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 have edges connecting each other: a **triangle**
Clustering coefficient

(a) Before new edges form.

(b) After new edges form.

- After a long period the graph on the right side (b) can be the result of the edges that were added to the original graph on the left side (a)
Clustering coefficient

![Graph (a) - Before new edges form.](image)

- Clustering coefficient of A = 1/6

![Graph (b) - After new edges form.](image)

- Clustering coefficient of A = 1/2

- After a long period the graph on the right side (b) can be the result of the edges that were added to the original graph on the left side (a).
Clustering coefficient

- Triadic closure is **natural**
  - The reason why B and C are more likely to become friends if they are friends of A is simply based on the opportunity for B and C to meet.

1. If A spends time with both B and C, then there is an increased chance that they will end up knowing each other and potentially becoming friends

2. The fact that each of B and C is friends with A (provided they are aware) gives them a basis for trusting each other that may be lacking in an arbitrary pair of unconnected people

3. If A is friends with B and C, then it becomes a source of latent stress in these relationships if B and C are not friends with each other.
Clustering coefficient

- Clustering coefficient is useful
  
  - For example, American teenage girls who have a low clustering coefficient in their network of friends are significantly more likely to commit suicide than those with high clustering coefficient [BM04]
Strength of acquaintances

› Information about jobs is intuitively relatively scarce
› Hearing about jobs from others means they have access to info you don’t

- A’s friends C, D, E connect her to a tightly-knit group of friends while B reaches into a different part of the network
  ⇒ A, C, D, E will be exposed to the same info
  ⇒ B offers A access to info she wouldn’t be able to access otherwise
Strength of acquaintances

- Bridges and local bridges

Edge \(<A,B>\) is a *bridge* meaning that its removal would place A and B in distinct connected components.

Bridges provide nodes with access to parts of the network that would be unreachable by other means.
Strength of acquaintances

- **Remember**: giant connected components tell us that bridges are presumably 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 multi-step paths that also span these worlds.

- Hence the previous graph is unlikely to standalone and he is more likely embedded into a larger graph.
Strength of acquaintances

- Let’s have a look at such a larger graph.

- The edge \( <A,B> \) is not the only path that connects its two endpoints
- There is a longer path through \( F, G \) and \( H \)
Strength of acquaintances

- A local bridge is an edge joining two nodes A and B in a graph if its endpoints A and B have no friends in common.

- In other words, deleting a local bridge between A and B would increase the distance between A and B to a value strictly larger than 2.

- We say that the span of a local bridge is the distance its endpoints would be from each other if the edge were deleted.
Strength of acquaintances

- Edge $<A,B>$ is a **local bridge** since the removal of its edge would increase the distance of $A$ and $B$ to 4 (its span is 4).

- Local bridges that have a large span play a role **similar to bridges**
- Local bridges are **more common** than bridges
Strong and Weak Ties
Strong and weak ties

Let’s classify edges depending on their strength into 2 classes:

- Strong ties
- Weak ties

For example, let assume nodes report on their close friends (strong ties) and acquaintances (weak ties).
Strong and weak ties

- Then we can label the edge of the previous graph (s=strong, w=weak)
Strong triadic closure

- We say that 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 (be it strong or weak) between B and C. We say that node A satisfies the Strong Triadic Closure property if it does not violate it.
Strong triadic closure

- All nodes in this figure satisfy the Strong Triadic Closure Property
Strong triadic closure

- Using **triadic closure** we can establish a **connection** between a purely **local** differentiation of edges and the **global** structural notion of bridges.

- **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**.

- Assuming the Strong Triadic Closure property and a sufficient number of strong ties, the local bridges in a network are necessarily **weak ties**.
Strong triadic closure

- Proof of the Claim (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
- Suppose A is involved in a local bridge, say with node B, that is a strong tie
- This is impossible:
  - 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 will call C
  - <B,C> cannot exist since <A,B> is a local bridge
  - This violates the Strong Triadic Closure property since <A,B> and <A,C> are both strong ties, the edge <B,C> must exist
**Strong triadic closure**

- **Proof of the Claim (by contradiction):**

![Diagram showing the concept of strong triadic closure with nodes A, B, and C, and edge S between A and B, with a note on the diagram stating: Strong Triadic Closure says the B-C edge must exist, but the definition of a local bridge says it cannot.](image-url)
Application to Large Scale
Who-talks-to-whom network

- Onnela et al. [OSH+07] studied the network maintained by a cell phone provider
  - Phone calls were essentially personal (not business phone calls)
  - Cell phone numbers were generally exchanged between people who know each other
  ⇒ Could be viewed as conversations occurring in a social network

- Results
  - It covered 20% of a national population
  - Nodes are cell phone users
  - Edges represent a phone call to each other in both directions over an 18 week observation period
  - A single giant connected component included 84% of the nodes
Generalization to large scale networks

- **Constraints** of our definitions:
  - An edge is either a weak tie or a strong tie
  - An edge is either a local bridge or not
- Can we have definitions that exhibit smoother gradation for larger scales?

- Previous example of the **who-talked-to-whom** network:
  - A strength can be a numerical quantity
    - total number of minutes spent on phone between the two ends of an edge in the previous example
  - Since a very small fraction of the edges are local bridges, let’s define the
    - Neighborhood overlap:

\[
\frac{\text{number of nodes who are neighbors of both } A \text{ and } B}{\text{number of nodes who are neighbors of at least one of } A \text{ or } B}
\]
Example of neighborhood overlap:

- There are 6 neighbors of A or F: B, C, D, E, G and J.
- There is only 1 of these that is common to both A’s neighbors and F’s neighbors: C
- Hence the neighborhood overlap is $1/6$
Generalization to large scale networks

- When does the neighborhood overlap nullify?
  - When the numerator is 0
  - This happens when the edge is a local bridge

⇒ Edges with low neighborhood overlap are “almost” local bridges

- \(<A,F>\) is closer to be a local bridge than \(<A,E>\) for example
Generalization to large scale networks

- We can derive interesting questions
  - How does the neighborhood overlap of an edge depend on its strength?
  - Neighborhood overlap should grow as strength grows

- The figure shows neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength

- The relationship between these quantities thus aligns well with the theoretical predictions
Generalization to large scale networks

- How these data can show that weak ties connect tightly-knit communities?
  - Onnela et al. [OSH+07] first deleted edges in descending order of their strength
  - The giant component shrank steadily
  - Then they re-started but by deleting edges in ascending order of their strength
  - The giant component shrank more rapidly
  - Its remnants broke abruptly once a critical number of weak ties were removed

⇒ This is consistent with a picture in which the weak ties provide the more crucial connective structure for holding together disparate communities
Tie strength on Facebook

- Facebook
  - *Friendship* is symmetric: if A is friend with B, then B is friend with A
  - Users post indirect messages, photos, videos, links that other users may see
    - A user’s *view* aggregates friends’ posts selected with a variant of EdgeRank
    - EdgeRank selected posts based on contents (text, link, media), related actions from others (e.g., comments, like, share) and staleness (of content and actions)
  - Users post direct public messages on other users wall
  - Users can send private messages to other users through (instant) messages
  - There are other mode of interactions, through walls, poke...
Tie strength on Facebook

- A link represents \textit{reciprocal (mutual) communication}, if the user both sent messages to the friend at the other end of the link, and also received messages from them during the observation period.

- A link represents a \textit{one-way communication} if the user sent one or more messages to the friend at the other end of the link (whether or not these messages were reciprocated).

- A link represents a \textit{maintained relationship} if the user followed information about the friend at the other end of the link, whether or not actual communication took place.
**Tie strength on Facebook**

- The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighborhood size for users on Facebook.
Tie strength on Facebook

- These three categories are not mutually exclusive [MBL+09]

- Even for users with large set of friends (~500)
  - they actually communicate with 10-20 and
  - they follow passively only 50 other users

- The set of users followed passively is interesting
  - It is a middle ground between weak and strong ties
  - It allows important news (e.g., new born baby) to disseminate very rapidly
Tie strength on Twitter

- **Twitter**
  - Online micro-blogging service
  - Users post *tweets* that are very short (140 characters) public messages
  - Users specify users they want to *follow*, i.e., from which they want to receive
  - Users can *direct* messages to another user (even though it is public, it is marked)
    \[\Rightarrow\] A *follower* of \( u \) is someone who follows \( u \)
    \[\Rightarrow\] A *followee* of \( u \) is someone that \( u \) follows

- **Followers** define *weak ties* of a social network
- **Directing messages** define a *stronger* kind of direct interaction
Tie strength on Twitter

- Twitter [HRW09]
  - A recent study considered:
    - A strong tie between u and v if u sent two direct messages to v during the observation
    - Users with 1000 followees have less than 50 strong ties
- The total number of a user’s strong ties (defined by multiple directed messages) as a function of the number of followees he or she has on Twitter
Tie strength on Twitter

- Understanding how social media have on the maintenance and use of social networks is complex
- But it seems that network of strong ties can still be sparse where weak ties abound
Closure and Holes
Neighborhood closure

- Access to edges that span different groups is not equally distributed across all nodes
  - Some nodes are at the interface between multiple groups
  - Others are placed at the middle of a single group

- What is the effect of this heterogeneity?
  - We can answer this question with a story about node experience in a network
Neighborhood closure

- The *Embeddedness* of an edge in the network is the number of common neighbors its two endpoints have.

- If two individuals are connected by an embedded edge then this makes it easier for them to trust one another.

- The presence of mutual friends puts the interactions between the two people “on display” in a social sense.

- In case of misbehavior of one of the two parties there is potential for social sanctions and reputational consequences from their mutual friend.

- Granovetter said “My mortification at cheating a friend of long standing may be substantial even when undiscovered. It may increase when a friend becomes aware of it. But it may become even more unbearable when our mutual friends uncover the deceit and tell one another.”
Neighborhood closure

- The contrast between densely-knit groups and boundary-spanning links is reflected in the different positions of nodes A and B in the underlying social network.

- A has a high clustering coefficient.

- All of A edges have significant embeddedness.

- Interactions between B, C and D are much more riskier.

- Moreover B is subject to potentially contradictory norms from its different groups.
Structural holes

- We discussed the advantages of a node based on its neighborhood closure
- But, what about advantages of being at the end of local bridges?
Structural holes

**Example:** Consider an organization [Bush45]

- **Nodes are managers**
  - Working on *common objectives*
  - Competing for career advancement

- **Edges are relations**
  - Employees are linked if they *know each other* and *talk to each other*
  - The edges do *not* represent necessarily *hierarchical relations*
Example: Consider an organization (con’t)

- B’s position offers advantages relative to A
  - B has access to information originating in multiple, non-interacting parts of the network
  ⇒ B is investing her energy efficiently by maintaining contacts across the organization
  - Standing at one edge of a local bridge can amplify creativity
  - B has a gatekeeper role by regulating accesses of C and D to her tightly knit groups
**Structural holes**

**Example:** Consider an organization (con’t)

- B is **powerful** as a gatekeeper
- Her **interest may not be aligned** with the one of the company
  - B may not want to lose the gatekeeping role
  - B may want to **prevent triangles** from forming around their local bridges
  - The organization may want to **accelerate the flow** of information between groups

- Empirical studies of managers in large corporations has correlated an individual’s success within a company to their access to local bridges \[86,87\]
- This network is **static** but how long these local bridges can last under triadic closure pressure is an **active research topic**?
Conclusion
Conclusion

- **Strength of ties is heterogeneous**
  - Intuitively, the strength translates into the frequency of interactions
  - The number of strong ties is generally low compared to the number of weak ties
  - **Weak ties** can be extremely **useful** in disseminating rare/important information in social networks

- **Triadic closure and clustering coefficient are natural concepts**
  - Tightly-knit regions induce implicit **trust** in social networks
  - Local bridges (in holes) often bring **importance** to its end points

- **Identifying** these tightly knit regions and holes
  - Is **useful** (to maximize information flow, detect powerful nodes)
  - But requires a **lightweight algorithm** to apply to large scale
References