Chapter 9

Social Networks

Distributed computing is applicable in various contexts. This lecture exemplarily studies one of these contexts, social networks, an area of study whose origins date back a century. To give you a first impression, consider Figure 9.1.

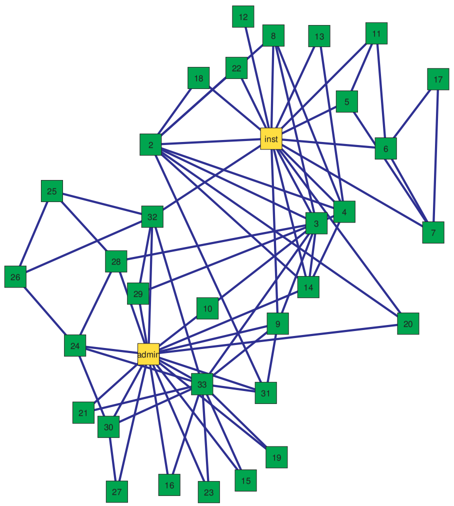

Figure 9.1: This graph shows the social relations between the members of a karate club, studied by anthropologist Wayne Zachary in the 1970s. Two people (nodes) stand out, the instructor and the administrator of the club, both happen to have many friends among club members. At some point, a dispute caused the club to split into two. Can you predict how the club partitioned? (If not, just search the Internet for Zachary and Karate.)

9.1 Small World Networks

Back in 1929, Frigyes Karinthy published a volume of short stories that postulated that the world was "shrinking" because human beings were connected more and more. Some claim that he was inspired by radio network pioneer Guglielmo Marconi's 1909 Nobel Prize speech. Despite physical distance, the growing density of human "networks" renders the actual social distance smaller and smaller. As a result, it is believed that any two individuals can be connected through at most five (or so) acquaintances, i.e., within six hops.

The topic was hot in the 1960s. For instance, in 1964, Marshall McLuhan coined the metaphor "Global Village". He wrote: "As electrically contracted, the globe is no more than a village". He argues that due to the almost instantaneous reaction times of new ("electric") technologies, each individual inevitably feels the consequences of his actions and thus automatically deeply participates in the global society. McLuhan understood what we now can directly observe – real and virtual world are moving together. He realized that the transmission medium, rather than the transmitted information is at the core of change, as expressed by his famous phrase "the medium is the message".

This idea has been followed ardently in the 1960s by several sociologists, first by Michael Gurevich, later by Stanley Milgram. Milgram wanted to know the average path length between two "random" humans, by using various experiments, generally using randomly chosen individuals from the US Midwest as starting points, and a stockbroker living in a suburb of Boston as target. The starting points were given name, address, occupation, plus some personal information about the target. They were asked to send a letter to the target. However, they were not allowed to *directly* send the letter, rather, they had to pass it to somebody they knew on first-name basis and that they thought to have a higher probability to know the target person. This process was repeated, until somebody knew the target person, and could deliver the letter. Shortly after starting the experiment, letters have been received. Most letters were lost during the process, but if they arrived, the average path length was about 5.5. The observation that the entire population is connected by short acquaintance chains got later popularized by the terms "six degrees of separation" and "small world".

Statisticians tried to explain Milgram's experiments, by essentially giving network models that allowed for short diameters, i.e., each node is connected to each other node by only a few hops. Until today there is a thriving research community in statistical physics that tries to understand network properties that allow for "small world" effects.

One of the keywords in this area are power-law graphs, networks were node degrees are distributed according to a power-law distribution, i.e., the number of nodes with degree δ is proportional to $\delta^{-\alpha}$, for some $\alpha > 1$. Such power-law graphs have been witnessed in many application areas, apart from social networks also in the web, or in Biology or Physics.

Obviously, two power-law graphs might look and behave completely differently, even if α and the number of edges is exactly the same.

One well-known model towards this end is the Watts-Strogatz model. Watts and Strogatz argued that social networks should be modeled by a combination of two networks: As the basis we take a network that has a large cluster coefficient . . .

Definition 9.1. *The cluster coefficient of a network is defined by the probability that two friends of a node are likely to be friends as well, averaged over all the nodes.*

. . . , then we augment such a graph with random links, every node for instance points to a constant number of other nodes, chosen uniformly at random. This augmentation represents acquaintances that connect nodes to parts of the network that would otherwise be far away.

Remarks:

- Without further information, knowing the cluster coefficient is of questionable value: Assume we arrange the nodes in a grid. Technically, if we connect each node to its four closest neighbors, the graph has cluster coefficient 0, since there are no triangles; if we instead connect each node with its eight closest neighbors, the cluster coefficient is 3/7. The cluster coefficient is quite different, even though both networks have similar characteristics.

This is interesting, but not enough to really understand what is going on. For Milgram's experiments to work, it is not sufficient to connect the nodes in a certain way. In addition, the nodes *themselves* need to know how to forward a message to one of their neighbors, even though they cannot know whether that neighbor is really closer to the target. In other words, nodes are not just following physical laws, but they make decisions themselves.

Let us consider an artificial network with nodes on a grid topology, plus some additional random links per node. In a quantitative study it was shown that the random links need a specific distance distribution to allow for efficient greedy routing. This distribution marks the sweet spot for any navigable network.

Definition 9.2 (Augmented Grid). *We take $n = m^2$ nodes $(i, j) \in V = \{1, \ldots, m\}^2$ that are identified with the lattice points on an $m \times m$ grid. We define the distance between two nodes (i, j) and (k, ℓ) as $d\big((i, j), (k, \ell)\big) = |k - i| + |\ell - j|$ as the distance between them on the $m \times m$ lattice. The network is modeled using a parameter $\alpha \geq 0$. Each node u has a directed edge to every lattice neighbor. These are the* local contacts *of a node. In addition, each node also has an additional random link (the* long-range contact*). For all u and v, the long-range contact of u points to node v with probability proportional to $d(u, v)^{-\alpha}$, i.e., with probability $d(u, v)^{-\alpha} / \sum_{w \in V \setminus \{u\}} d(u, w)^{-\alpha}$. Figure 9.2 illustrates the model.*

Remarks:

- The network model has the following geographic interpretation: nodes (individuals) live on a grid and know their neighbors on the grid. Further, each node has some additional acquaintances throughout the network.

- The parameter α controls how the additional neighbors are distributed across the grid. If $\alpha = 0$, long-range contacts are chosen uniformly at random (as in the Watts-Strogatz model). As α increases, long-range contacts become shorter on average. In the extreme case, if $\alpha \to \infty$, all long-range contacts are to immediate neighbors on the grid.

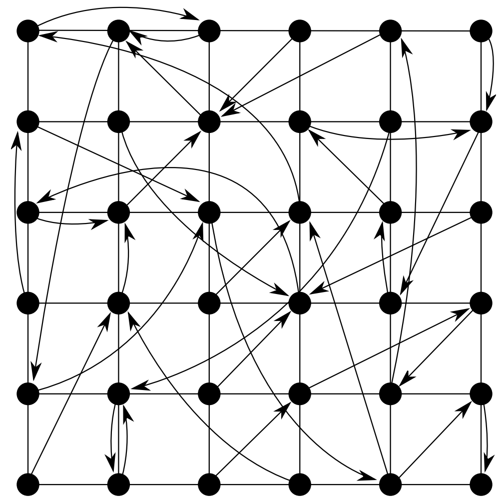

Figure 9.2: Augmented grid with $m = 6$

- It can be shown that as long as $\alpha \leq 2$, the diameter of the resulting graph is polylogarithmic in n (polynomial in $\log n$) with high probability. In particular, if the long-range contacts are chosen uniformly at random ($\alpha = 0$), the diameter is $\mathcal{O}(\log n)$.

Since the augmented grid contains random links, we do not know anything for sure about how the random links are distributed. In theory, all links could point to the same node! However, this is almost certainly not the case. Formally this is captured by the term *with high probability.*

Definition 9.3 (With High Probability). *Some probabilistic event is said to occur* with high probability (w.h.p.)*, if it happens with a probability $p \geq 1 - 1/n^c$, where c is a constant. The constant c may be chosen arbitrarily, but it is considered constant with respect to Big-O notation.*

Remarks:

- For instance, a running time bound of $c \log n$ or $e^{c!} \log n + 5000c$ with probability at least $1 - 1/n^c$ would be $\mathcal{O}(\log n)$ w.h.p., but a running time of n^c would not be $\mathcal{O}(n)$ w.h.p. since c might also be 50.

- This definition is very powerful, as any polynomial (in n) number of statements that hold w.h.p. also holds w.h.p. at the same time, regardless of any dependencies between random variables!

Theorem 9.4. *The diameter of the augmented grid with $\alpha = 0$ is $\mathcal{O}(\log n)$ with high probability.*

Proof Sketch. For simplicity, we will only show that we can reach a target node t starting from some source node s. However, it can be shown that (essentially) each of the intermediate claims holds with high probability, which then by means of the union bound yields that *all* of the claims hold simultaneously with high probability for *all* pairs of nodes (see exercises).

Let N_s be the $\lceil \log n \rceil$-hop neighborhood of source s on the grid, containing $\Omega(\log^2 n)$ nodes. Each of the nodes in N_s has a random link, probably leading to distant parts of the graph. As long as we have reached only $o(n)$ nodes, any new random link will with probability $1 - o(1)$ lead to a node for which none of its grid neighbors has been visited yet. Thus, in expectation we find almost $|N_s|$ new nodes whose neighbors are "fresh". Using their grid links, we will reach $(4 - o(1))|N_s|$ more nodes within one more hop. If bad luck strikes, it could still happen that many of these links lead to a few nodes, already visited nodes, or nodes that are very close to each other. But that is very unlikely, as we have lots of random choices! Indeed, it can be shown that not only in expectation, but with high probability $(5 - o(1))|N_s|$ many nodes are reached this way (see exercises).

Because all the new nodes have (so far unused) random links, we can repeat this reasoning inductively, implying that the number of nodes grows by (at least) a constant factor for every two hops. Thus, after $\mathcal{O}(\log n)$ hops, we will have reached $n/\log n$ nodes (which is still small compared to n). Finally, consider the expected number of links from these nodes that enter the $(\log n)$-neighborhood of some target node t with respect to the grid. Since this neighborhood consists of $\Omega(\log^2 n)$ nodes, in expectation $\Omega(\log n)$ links come close enough to target t. This is large enough to almost guarantee that this happens (see exercises). Summing everything up, we still used merely $\mathcal{O}(\log n)$ hops in total to get from s to t.

□

This shows that for $\alpha = 0$ (and in fact for all $\alpha \leq 2$), the resulting network has a small diameter. Recall however that we also wanted the network to be navigable. For this, we consider a simple greedy routing strategy (Algorithm 39).

Algorithm 39 Greedy Routing

while not at destination **do**
  go to a neighbor which is closest to destination (considering grid distance only)
end while

Lemma 9.5. *In the augmented grid, Algorithm 39 finds a routing path of length at most $2(m-1) \in O(\sqrt{n})$.*

Proof. Because of the grid, there is always a neighbor which is closer to the destination. Since with each hop we reduce the distance to the target at least by one in one of the two grid dimensions, we will reach the destination within $2(m-1)$ steps. □

This is not really what Milgram's experiment promises. We want to know how much the additional random links speed up the process. To this end, we first need to understand how likely it is that the random link of node u points to node v, in terms of their grid distance $d(u,v)$, the number of nodes n, and the constant parameter α.

Lemma 9.6. *Node u's random link points to a node v with probability*

- $\Theta(1/(d(u,v)^\alpha m^{2-\alpha}))$ *if* $\alpha < 2$.
- $\Theta(1/(d(u,v)^2 \log n))$ *if* $\alpha = 2$,
- $\Theta(1/d(u,v)^\alpha)$ *if* $\alpha > 2$.

Moreover, if $\alpha > 2$, the probability to see a link of length at least d is in $\Theta(1/d^{\alpha-2})$.

Proof. For a constant $\alpha \neq 2$, we have that

$$\sum_{w \in V \setminus \{u\}} \frac{1}{d(u,w)^\alpha} \in \sum_{r=1}^{m} \frac{\Theta(r)}{r^\alpha} = \Theta\left(\int_{r=1}^{m} \frac{1}{r^{\alpha-1}}\, dr\right) = \Theta\left(\left[\frac{r^{2-\alpha}}{2-\alpha}\right]_1^m\right).$$

If $\alpha < 2$, this gives $\Theta(m^{2-\alpha})$, if $\alpha > 2$, it is in $\Theta(1)$. If $\alpha = 2$, we get

$$\sum_{w \in V \setminus \{u\}} \frac{1}{d(u,w)^\alpha} \in \sum_{r=1}^{m} \frac{\Theta(r)}{r^2} = \Theta(1) \cdot \sum_{r=1}^{m} \frac{1}{r} = \Theta(\log m) = \Theta(\log n).$$

Multiplying with $d(u,v)^\alpha$ yields the first three bounds. For the last statement, compute

$$\sum_{\substack{v \in V \\ d(u,v) \geq d}} \Theta(1/d(u,v)^\alpha) = \Theta\left(\int_{r=d}^{m} \frac{r}{r^\alpha}\, dr\right) = \Theta\left(\left[\frac{r^{2-\alpha}}{2-\alpha}\right]_d^m\right) = \Theta(1/d^{\alpha-2}).$$

□

Remarks:

- If $\alpha > 2$, according to the lemma, the probability to see a random link of length at least $d = m^{1/(\alpha-1)}$ is $\Theta(1/d^{\alpha-2}) = \Theta(1/m^{(\alpha-2)/(\alpha-1)})$. In expectation we have to take $\Theta(m^{(\alpha-2)/(\alpha-1)})$ hops until we see a random link of length at least d. When just following links of length less than d, it takes more than $m/d = m/m^{1/(\alpha-1)} = m^{(\alpha-2)/(\alpha-1)}$ hops. In other words, in expectation, either way we need at least $m^{(\alpha-2)/(\alpha-1)} = m^{\Omega(1)}$ hops to the destination.

- If $\alpha < 2$, there is a (slightly more complicated) argument. First we draw a border around the nodes in distance $m^{(2-\alpha)/3}$ to the target. Within this border there are about $m^{2(2-\alpha)/3}$ many nodes in the target area. Assume that the source is outside the target area. Starting at the source, the probability to find a random link that leads directly inside the target area is according to the lemma at most $m^{2(2-\alpha)/3} \cdot \Theta(1/m^{2-\alpha})) = \Theta(1/m^{(2-\alpha)/3})$. In other words, until we find a random link that leads into the target area, in expectation, we have to do $\Theta(m^{(2-\alpha)/3})$ hops. This is too slow, and our greedy strategy is probably faster, as thanks to having $\alpha < 2$ there are many long-range links. However, it means that we will probably enter the border of the target area on a regular grid link. Once inside the target area, again the probability of short-cutting our trip by a random long-range link is $\Theta(1/m^{(2-\alpha)/3})$, so we probably just follow grid links, $m^{(2-\alpha)/3} = m^{\Omega(1)}$ many of them.

- In summary, if $\alpha \neq 2$, our greedy routing algorithm takes $m^{\Omega(1)} = n^{\Omega(1)}$ expected hops to reach the destination. This is polynomial in the number of nodes n, and the social network can hardly be called a "small world".

- Maybe we can get a polylogarithmic bound on n if we set $\alpha = 2$?

Definition 9.7 (Phase)**.** *Consider routing from source s to target t and assume that we are at some intermediate node w. We say that we are in phase j at node w if the lattice distance $d(w,t)$ to the target node t is between $2^j < d(w,t) \leq 2^{j+1}$.*

Remarks:

- Enumerating the phases in decreasing order is useful, as notation becomes less cumbersome.

- There are $\lceil \log m \rceil \in O(\log n)$ phases.

Lemma 9.8. *Assume that we are in phase j at node w when routing from s to t. The probability for getting (at least) to phase $j-1$ in one step is at least $\Omega(1/\log n)$.*

Proof. Let B_j be the set of nodes x with $d(x,t) \leq 2^j$. We get from phase j to (at least) phase $j-1$ if the long-range contact of node w points to some node in B_j. Note that we always make progress while following the greedy routing path. Therefore, we have not seen node w before and the long-range contact of w points to a random node that is independent of anything seen on the path from s to w.

For all nodes $x \in B_j$, we have $d(w,x) \leq d(w,t) + d(x,t) \leq 2^{j+1} + 2^j < 2^{j+2}$. Hence, for each node $x \in B_j$, the probability that the long-range contact of w points to x is $\Omega(1/2^{2j+4}\log n)$. Further, the number of nodes in B_j is at least $(2^j)^2/2 = 2^{2j-1}$. Hence, the probability that some node in B_j is the long range contact of w is at least

$$\Omega\left(|B_j| \cdot \frac{1}{2^{2j+4}\log n}\right) = \Omega\left(\frac{2^{2j-1}}{2^{2j+4}\log n}\right) = \Omega\left(\frac{1}{\log n}\right). \qquad \square$$

Theorem 9.9. *Consider the greedy routing path from a node s to a node t on an augmented grid with parameter $\alpha = 2$. The expected length of the path is $\mathcal{O}(\log^2 n)$.*

Proof. We already observed that the total number of phases is $\mathcal{O}(\log n)$ (the distance to the target is halved when we go from phase j to phase $j-1$). At each point during the routing process, the probability of proceeding to the next phase is at least $\Omega(1/\log n)$. Let X_j be the number of steps in phase j. Because

the probability for ending the phase is $\Omega(1/\log n)$ in each step, in expectation we need $\mathcal{O}(\log n)$ steps to proceed to the next phase, i.e., $\mathbb{E}[X_j] \in O(\log n)$. Let $X = \sum_j X_j$ be the total number of steps of the routing process. By linearity of expectation, we have

$$\mathbb{E}[X] = \sum_j \mathbb{E}[X_j] \in O(\log^2 n). \qquad \square$$

Remarks:

- One can show that the $\mathcal{O}(\log^2 n)$ result also holds w.h.p.

- In real world social networks, the parameter α was evaluated experimentally. The assumption is that you are connected to the geographically closest nodes, and then have some random long-range contacts. For Facebook grandpa LiveJournal it was shown that α is not really 2, but rather around 1.25.

9.2 Propagation Studies

In networks, nodes may influence each other's behavior and decisions. There are many applications where nodes influence their neighbors, e.g., they may impact their opinions, or they may bias what products they buy, or they may pass on a disease.

On a beach (modeled as a line segment), it is best to place an ice cream stand right in the middle of the segment, because you will be able to "control" the beach most easily. What about the second stand, where should it settle? The answer generally depends on the model, but assuming that people will buy ice cream from the stand that is closer, it should go right next to the first stand.

Rumors can spread surprisingly fast through social networks. Traditionally this happens by word of mouth, but with the emergence of the Internet and its possibilities new ways of rumor propagation are available. People write email, use instant messengers or publish their thoughts in a blog. Many factors influence the dissemination of rumors. It is especially important where in a network a rumor is initiated and how convincing it is. Furthermore the underlying network structure decides how fast the information can spread and how many people are reached. More generally, we can speak of diffusion of information in networks. The analysis of these diffusion processes can be useful for viral marketing, e.g., to target a few influential people to initiate marketing campaigns. A company may wish to distribute the rumor of a new product via the most influential individuals in popular social networks such as Facebook. A second company might want to introduce a competing product and has hence to select where to seed the information to be disseminated. Rumor spreading is quite similar to our ice cream stand problem.

More formally, we may study propagation problems in graphs. Given a graph, and two players. Let the first player choose a seed node u_1; afterwards let the second player choose a seed node u_2, with $u_2 \neq u_1$. The goal of the game is to maximize the number of nodes that are closer to one's own seed node.

In many graphs it is an advantage to choose first. In a star graph for instance the first player can choose the center node of the star, controlling all but one node. In some other graphs, the second player can at least score even. But is there a graph where the second player has an advantage?

Theorem 9.10. *In a two player rumor game where both players select one node to initiate their rumor in the graph, the first player does not always win.*

Proof. See Figure 9.3 for an example where the second player will always win, regardless of the decision the first player. If the first player chooses the node x_0 in the center, the second player can select x_1. Choice x_1 will be outwitted by x_2, and x_2 itself can be answered by z_1. All other strategies are either symmetric, or even less promising for the first player. □

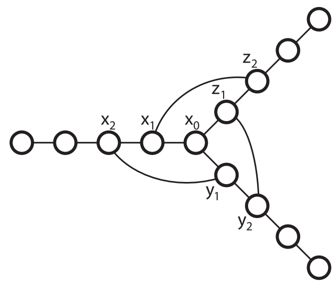

Figure 9.3: Counter example.

Chapter Notes

A simple form of a social network is the famous stable marriage problem [DS62] in which a stable matching bipartite graph has to be found. There exists a great many of variations which are based on this initial problem, e.g., [KC82, KMV94, EO06, FKPS10, Hoe11]. Social networks like Facebook, Twitter and others have grown very fast in the last years and hence spurred interest to research them. How users influence other users has been studied both from a theoretical point of view [KKT03] and in practice [CHBG10]. The structure of these networks can be measured and studied [MMG+07]. More than half of the users in social networks share more information than they expect to [LGKM11].

The small world phenomenon that we presented in this chapter is analyzed by Kleinberg [Kle00]. A general overview is in [DJ10].

This chapter has been written in collaboration with Michael Kuhn.

Bibliography

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and P. Krishna Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*, 2010.

[DJ10] Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[DS62] D. Gale and L.S. Shapley. College Admission and the Stability of Marriage. *American Mathematical Monthly*, 69(1):9–15, 1962.

[EO06] Federico Echenique and Jorge Oviedo. A theory of stability in many-to-many matching markets. *Theoretical Economics*, 1(2):233–273, 2006.

[FKPS10] Patrik Floréen, Petteri Kaski, Valentin Polishchuk, and Jukka Suomela. Almost Stable Matchings by Truncating the Gale-Shapley Algorithm. *Algorithmica*, 58(1):102–118, 2010.

[Hoe11] Martin Hoefer. Local Matching Dynamics in Social Networks. *Automata Languages and Programming*, pages 113–124, 2011.

[Kar29] Frigyes Karinthy. Chain-Links, 1929.

[KC82] Alexander S. Kelso and Vincent P. Crawford. Job Matching, Coalition Formation, and Gross Substitutes. *Econometrica*, 50(6):1483–1504, 1982.

[KKT03] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[Kle00] Jon M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC*, 2000.

[KMV94] Samir Khuller, Stephen G. Mitchell, and Vijay V. Vazirani. On-line algorithms for weighted bipartite matching and stable marriages. *Theoretical Computer Science*, 127:255–267, May 1994.

[LGKM11] Yabing Liu, Krishna P. Gummadi, Balanchander Krishnamurthy, and Alan Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, Berlin, Germany, November 2011.

[McL64] Marshall McLuhan. *Understanding media: The extensions of man.* McGraw-Hill, New York, 1964.

[Mil67] Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967.

[MMG+07] Alan Mislove, Massimiliano Marcon, P. Krishna Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Internet Measurement Comference*, 2007.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of /'small-world/' networks. *Nature*, 393(6684):440–442, jun 1998.

[Zac77] W W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.