1 Advanced Spreading Processes

In this lecture, we will explore more deeply (i) how network structure can drive more complicated and more realistic dynamics in the evolution of epidemics, and (ii) other models of spreading processes, such as those that represent how social adoption spreads.

1.1 Epidemic models with structure

Simple epidemic models running on uncomplicated networks produce simple epidemic dynamics. These tend to look like

1. a single exponential growth phase at the beginning of the spread,
2. a single peak, and
3. either a steady state (SIS) or a single fall back to the baseline (SIR).

Varying the degree structure, or even the density of triangles, or even introducing simple forms of community structure, does not change this pattern much. It may broaden or sharpen the single peak, but the single-peak pattern remains. However, when we look at the dynamics of real epidemics, we see far more complicated behaviors, and shown in Fig. 1 below.

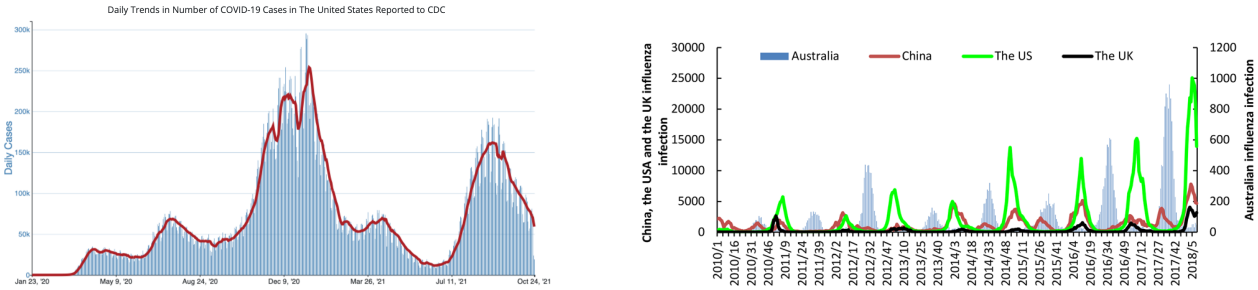


Figure 1: (left) COVID-19 case time series for the U.S. from January 2020 to October 2021, as recorded by the U.S. Center for Disease Control, showing five distinct "waves" of cases. Reproduced from the CDC COVID Tracker. (right) Influenza (all strains) case time series for four countries from 2010–2018, showing annual waves of different sizes, which oscillate between northern and southern hemisphere countries. Reproduced from Zhang et al. (2019).

There are many specific reasons why real epidemics have more irregular shapes or even run in wave-like dynamics, with multiple peaks, and different rates of increase or decrease. Recall that the dynamics of an epidemic depend on both (i) the biology of transmission and recovery (modeled by β and γ) and (ii) the shape of the exposure network G, where we assume that edges are the mechanisms of transmission.

Hence, from the perspective of network epidemiology, irregular epidemic dynamics must be driven by one or the other of these two parts of the model.

- Changes in either the basic transmissibility of the pathogen β or the recovery rate γ can increase or decrease the rate of spread, even with a fixed exposure graph.

  These changes could be driven by evolution of the pathogen itself, as in the shift from the wildtype SARS-CoV-2 virus to the alpha variant and then to the delta variant over the first 18 months of the COVID-19 pandemic, or in the way the environment interacts with transmission, e.g., because of changes in humidity or temperature, or by changes in medical interventions, such as improved treatments.

- Changes in the network of exposures G, driven by the dynamics of the social contact network.

  These changes can be caused by any number of things that shape who contacts with whom in our social world. A few examples with known relevant to disease spread include (i) time-varying adoption rates of non-pharmaceutical intervention (NPI) behaviors like social distancing, mask wearing, hand washing, self-quarantining, economic closures, etc., (ii) environmental influences, such as seasonal temperature changes, and (iii) even by the opening up of new venues for social contact, such as in opening a new transportation link between two places or opening up a new concert venue.

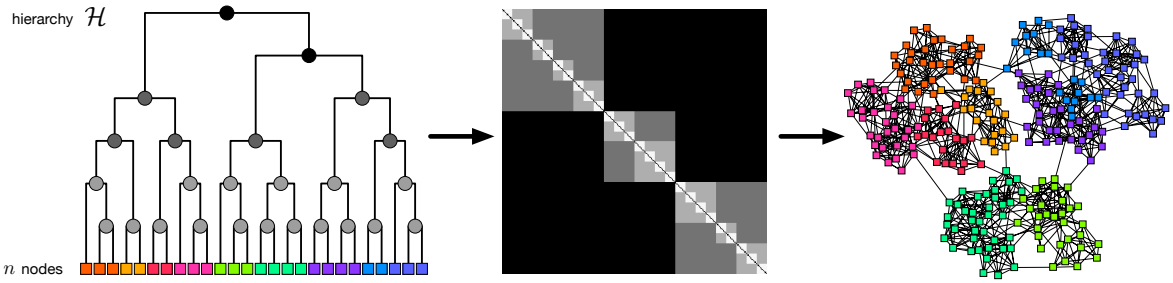
Here, we will explore the second category, in which the structure of the network shapes the dynamics of the spread itself to ensure that the epidemic doesn't happen everywhere at the same time.

1.1.1 Hierarchically structured populations

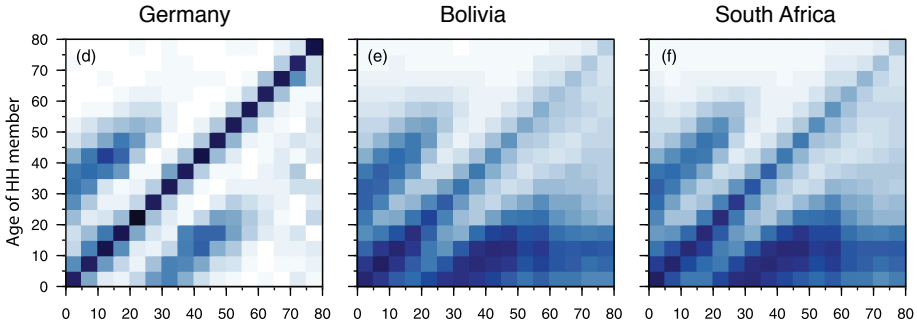
Real social networks are very, very large. At this scale, there's plenty of room for multiple levels of structural patterns, and we can model these via "hierarchical" structure, in which nodes tend to divide into groups that further subdivide into groups of groups, and so forth over multiple scales—nations, regions, cities, towns, neighborhoods, and households, or loose associations and the many smaller, closer groups they are composed of. This kind of hierarchical structure can appear for many different reasons, and is a good general way to represent multi-scale structural patterns in networks.

We typically represent the hierarchical organization of a network using a tree or a dendrogram $\mathcal{H}$, in which the nodes of the network are the leaves of the tree, and the internal branches show how we divide things into progressively smaller groups, as we move down the hierarchy from the whole system to its components.[1] The "group" that two nodes i, j belong to is the one defined by their lowest common ancestor in the hierarchy. Hence, the higher up in the hierarchy the lowest common ancestor, the more leaf nodes are below it and the larger the group it represents.

[1] Binary trees are a convenient choice for hierarchical structure, but any type of tree suffices.



We can turn a hierarchy of groups $\mathcal{H}$ into a network model by defining the probability that two nodes i, j are connected to be proportional to their "relatedness" in the hierarchy, where nodes with a smaller distance on the tree (lower common ancestor) are more related.[2] This assumption produces assortative "nested" communities, because the within-group density of edges increases as we consider smaller groups (moving down the hierarchy; see figure below). This pattern models the idea that you are more likely to interact with someone who lives in your neighborhood, or takes a class with you, or belongs to the same social organizations, than with someone who lives far away, or has divergent social interests.



One particular kind of large-scale population structure that we can model as a kind of hierarchy is an *age-structured population*, where most people interact with others who are within about 5 years of their own age, and they interact at slightly lower rates for people within 5–10 years of their age, etc.,[3] which is like the stylized "ordered" structure (or linear group hierarchy) we saw in Lecture 7, plus a family-based interaction pattern in the off-diagonal region: children interact at high rates

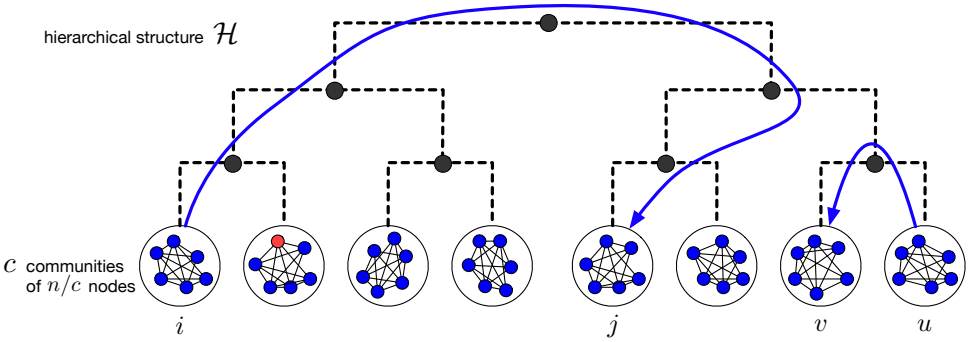
[2] See Clauset, Moore & Newman, "Hierarchical structure and the prediction of missing links in networks." *Nature* **483**, 98 (2008) for more details of how this model works.

[3] Figures from Prem, Cook & Jit, *PLoS Comp. Bio.* **13**, e1005697 (2017)

with individuals about 25–35 years (and about 60 years) older than them. On this kind of network, how would an epidemic spread? What prevents the epidemic from being everywhere at the same time?

1.1.2 Epidemics on hierarchical networks

The effect of hierarchical structure on epidemics is to prevent the epidemic from happening everywhere at once, and instead allowing it to repeatedly find naive subpopulations that together produce the ups and downs of realistic case counts. But rather than model the hierarchical network explicitly, we can capture its implicit structure using a hierarchical metapopulation simulation.[4]



To begin, we create c communities, and populate each with n/c nodes. Each of these communities will run its own SI-X compartment model, meaning that we assume each community is itself well mixed, with a common choice of β and γ parameters. Initially, all nodes are in the S state, and we infect a single node $x_r = I$, chosen at random. Then, with probability p, at each step of the model, each node relocates from its current community i to another community j chosen with probability $q_{ij} \propto \mathrm{e}^{-\lambda d_{ij}}$, where d_{ij} is the tree distance between i and j.[5] The simulation then runs until the epidemic ends, or for as long as we like.

In this model, the parameter p acts like a global mixing rate, as it determines the frequency at which nodes relocate to different communities. When p is small, the community-level epidemics

`http://dx.doi.org/10.1371/journal.pcbi.1005697`.

[4]This model was first described and explored in Watts et al., "Multiscale, resurgent epidemics in a hierarchical metapopulation model." *Proc. Natl. Acad. Sci. USA* **102**, 11157 (2005).

[5]We can efficiently calculate a distance d_{ij} by assigning a binary string of length $\log_2 c$ to each community (and hence to all nodes in that community), where each string encodes the sequence of left (0) or right (1) moves on the hierarchy, traversing from the root down to that community. Then, the tree distance d_{ij} for communities i and j is simply the edit distance or L1 norm of the two strings. In the figure here, community i would have label 000, j would have label 100, u label 111, and v label 110. If the hierarchy is a b-ary, meaning each branch has b children, then we use a b-ary string of length $\log_b c$.

remain relatively independent, but once p is above some threshold, a local epidemic will tend to export infected individuals to other communities, allowing it to spread globally. In contrast, λ sets the scale of that spread by tuning the typical distance over which a node will tend to relocate. When λ is large, nodes relocate to more nearby communities, and when λ is small, nodes tend to travel further. In the figure above, we have a depth 3 binary tree as the hierarchy on a set of 8 communities, and two relocations are shown as blue arrows: one long jump, and one short jump.

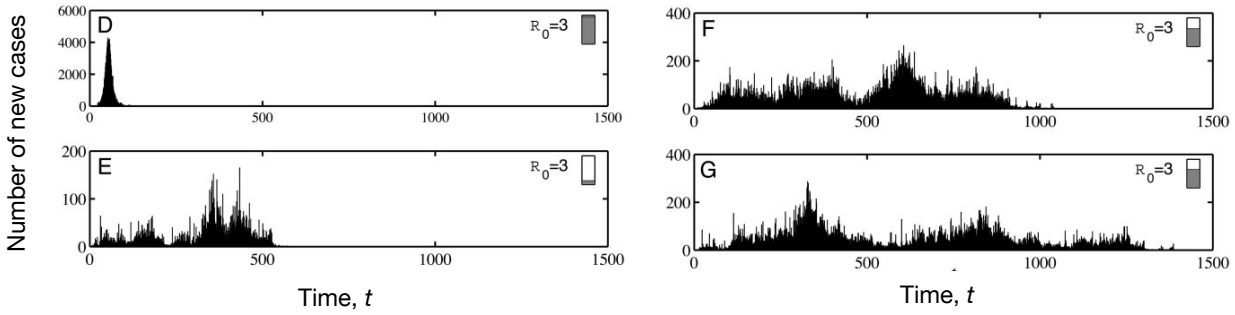


Figure 2: Simulated new case counts $\Delta I(t)$ over time, for the hierarchical metapopulation model with $n = 102,400$ nodes and $\beta/\gamma = 0.3/0.1 = 3$. (D) No hierarchy, only a single large community, showing classic SIR dynamics. (E-G) Hierarchy, with a $b = 4$ branching ratio in a tree with depth $\ell = 5$ and $n/c = 100$ nodes in each of the $c = 1024$ communities, showing complicated dynamics of repeated resurgence. Adapted from Watts et al. (2005).

The epidemic dynamics produced by this model are far more realistic than the classic SIR model (see Fig. 2), reflecting the fact that the spreading process continually seeds epidemics in naive (still susceptible) subpopulations of different sizes. Each "resurgence" of the case numbers corresponds to such a new local epidemic starting up, and if we marked along the time axis the times at which an infected individual relocated to a previously fully susceptible community, these marks would line up with the upward swings of the curve. Within the model, the parameters p and λ ultimately shape the speed and scale at which the epidemic spreads across communities, and tune how densely or sparsely spaced out these new local epidemics are.

1.1.3 Real-world hierarchical structure

Large-scale transportation networks, such as roads, trains, and airplanes, have an explicit hierarchical structure because humans tend to engineer such systems in this way.[6] Roads are arranged in

[6]Of course, scientists have all sorts of explanations for why this is the case. Some have to do with the clumpiness of humans in space, which may follow a fractal-like pattern; another explanation says that self-similar clumpiness just follows the fractal-like structure of the natural environment. Other explanations argue that hierarchies provide certain kinds of efficiencies and robustnesses to perturbations.

major and minor highways, main arteries, all the way down to side streets, and the larger the road, the more traffic it carries and the more distance locations it tends to connect. Train, shipping, and air travel networks all have similar structure.[7]

For understanding the global dynamics of pandemics, like influenza and COVID-19, we must focus on transportation at the global scale, which means mainly air travel. Here, nodes are airports, and two airports are connected by a directed edge with weight w_{ij} equal to the mean number of daily passengers that fly from airport i to airport j, and this network plays the role of the hierarchy $\mathcal{H}$ in our metapopulation model, by governing the way individuals relocate from one community (city with an airport) to another.

The structure of the global air travel network is well-studied, and exhibits many of the same statistical patterns seen in other real-world networks: a heavy-tailed degree distribution, community structure, and triangles. In addition, the network is spatial: nodes (airports) are embedded in a 2-dimensional space, the surface of the Earth. Empirically, the probability that two airports are connected $\Pr(i \to j)$ decreases exponentially with their spatial separation d_{ij}, modulo some complexities due to the spatial arrangement of landmasses on Earth's surface. We can see some of these features in the figure above, with $n = 4069$ airports worldwide and the corresponding edges. The largest-degree nodes, like London Heathrow (UK), Guangzhou China, and Atlanta GA are also the most central, with the most flights passing through them and having the smallest number of hops from them to any other airport in the system. Hence, traveling from an airport i to some other airport j, is likely to go "up" in the hierarchy of airports, toward the more major hubs, and then back down to more local airports for the final step. This pattern is very much like navigating the road network, where crossing larger distances in space typically requires using "larger" roads.[8]

[7] For an excellent review of the shape of real-world transportation networks, see Barthlémy, "Spatial networks." *Physics Reports* **499**, 1–101 (2011).

[8] These two figures reproduced from Brockmann and Helbing, "The Hidden Geometry of Complex, Network-Driven Contagion Phenomena." *Science* **342**, 1337 (2013).

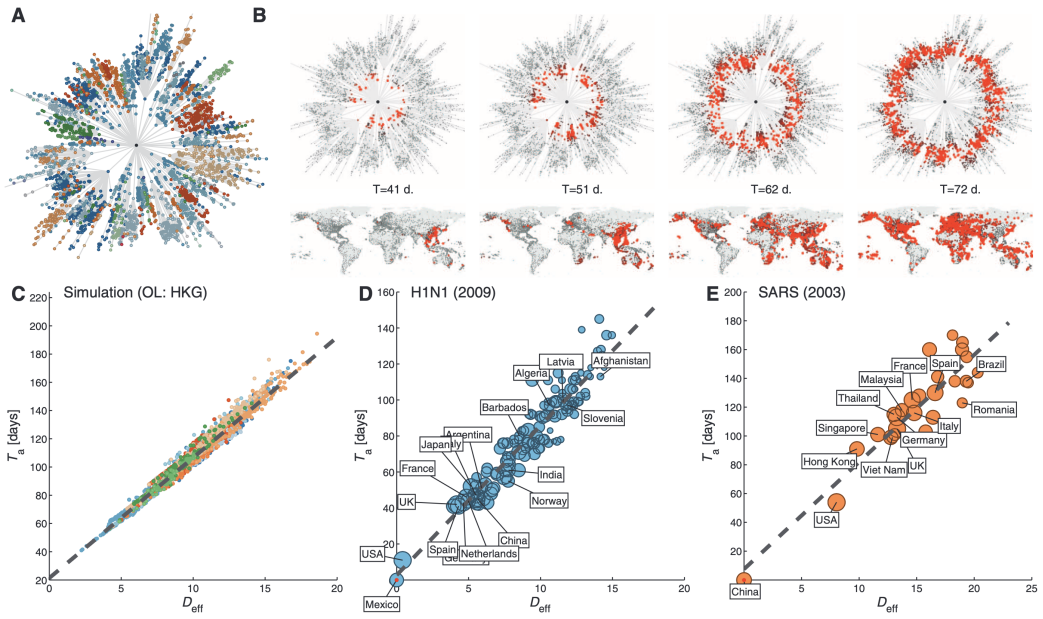


Figure 3: (A) The shortest-path tree from Hong Kong airport (HKG), where radial positions are correspond to effective distances D, and node color follows the network figure above. (B) Simulated propagation waves of infection, spreading outward from HKG. (C) Simulated "arrival times" of the epidemic as a function of total effective distance from HKG. (D, E) Empirical arrival times of two pandemics (H1N1 in 2009 and SARS-CoV-1 in 2003) as a function of total effective distance from their points of origin; the excellent agreement between simulation and data here suggests that the global spread of these two pandemics was largely driven by the airport network's structure. Reproduced from Brockmann & Helbing (2013).

The insight that the air travel network acts as the mechanism of relocations of individuals among "communities" at the global scale suggests that we can just replace the abstract hierarchical structure $\mathcal{H}$ in our metapopulation model with the airport network itself. This change yields a more realistic model of the global spread of a pandemic, and allows us to examine how proximity within the network can increase the risk of disease importation. It does not tell us how epidemic dynamics below this level of representation might feed into or be driven by the dynamics at this global scale, but one could imagine replacing the SI-X model within each community with a more fine-grained network model that captures the particular age or population structure of that community.

In a 2013 paper, Brockmann and Helbing constructed precisely this kind of air travel-based hierarchical metapopulation model. Their key insight was that some edges in the airport network carry far more passengers than others, and hence that link is more likely to transfer an infected individual from one community to the other. To parameterize their hierarchy, they first computed the fractional average flow of daily passengers $w_{ij} = f_{ij} / \sum_j f_{ij}$, where f_{ij} is the mean number of daily passengers f_{ij} flowing from node i to node j. These flows could be used to determine a

particular individual's trajectory through the network, e.g., by interpreting these weights was probabilities in a first-order Markov chain. But, to understand how the structure of the flows overall shape an epidemic, they instead used them to calculate a pairwise distance matrix D, where D_{ij} is the total "effective distance" of the edges in a geodesic path from $i \to j$, where the effective distance of an edge is defined as $d_{ij} = (1 - \log w_{ij})$ (do you see why this transformation is reasonable?).

From this perspective, we now have an effective hierarchy $\mathcal{H}$: an infected individual is more likely to transit between airports with smaller effective distance D, implying that these airports are closer together in the hierarchy. Running a standard SIR epidemic over this structurally-parameterized hierarchical metapopulation model produces remarkably realistic simulations of how real pandemics have spread, such as the H1N1 flu in 2009 and the SARS-CoV-1 in 2003 (see Fig. 3), with the arrival time of the epidemic correlating very strongly with the effective distance between the source and the arrival point within the implied airport network hierarchy.

1.2 Social adoption models and cascades

In a social spreading (aka, social "diffusion") process, we consider the question of how individuals influence each other to adopt some behavior or carry out some action, and how that behavior then percolates across a network as a result of that social connection-mediated influence. This context contrasts with the epidemiological setting, where the inherent dynamics of biological infection make a change in state from "susceptible" to "infected" are negative (because almost always, the thing that's spreading is exploiting the host's resources for outcomes that harm the host). Social spreading is more variable: it can be positive, as in the case of useful information spreading, such as job ads or health information, or negative, as in the case of gossip or conspiracy theories, or more ambiguous, as in the case of many memes and other social content online.

1.2.1 Two standard models

The two most common models of social adoption are the *independent cascade* (IC) model and the *linear threshold* (LT) model, although there are many variations and elaborations of each.[9] Both use a set of node states $\Gamma = \{I, A\}$ to indicate nodes as being either *inactive* or *active*.

Independent Cascades. The IC model is like a one-shot SI model, where each edge gets exactly one chance to apply the edge-update rule from Lecture 9; if it fails, that edge stays "quiet" forever. This behavior models the idea that if i attempts but fails to influence a neighbor j, then j can never by convinced by i.

[9] A very early example of using spreading processes to model social adoption is Goffman and Newill, "Generalizations of epidemic theory: An application to the transmission of ideas." *Nature* **4955**, 225–228 (1964), which modeled the spread of scientific ideas within a population using an SIR model.

For formally, the IC model works as follows.[10] given a directed or undirected network $G = (V, E)$, and a set of states $\Gamma = \{I, A\}$ (inactive and active), and we initialize some subset of the nodes to be active.[11]

Time then proceeds in discrete steps according to the following update rule. When a node u first becomes active $x_u = A$ at time t, it has a single chance in time step t to activate each of currently inactive neighbor v, which occurs with probability p_{uv} and is independent of any previous activations. If v has multiple newly active neighbors, each has an independent chance to activate v. If v is activated, its state updates $x_v = A$ for the next time step. Time proceeds until no more new activations occur. The edge values p_{uv} are parameters of the model, but are often set to a constant probability of transmission $p_{uv} = p$. In the special case where $p = 1$, the IC cascade traces out the single-source shortest path trees (aka, breadth-first trees) from the seed nodes.

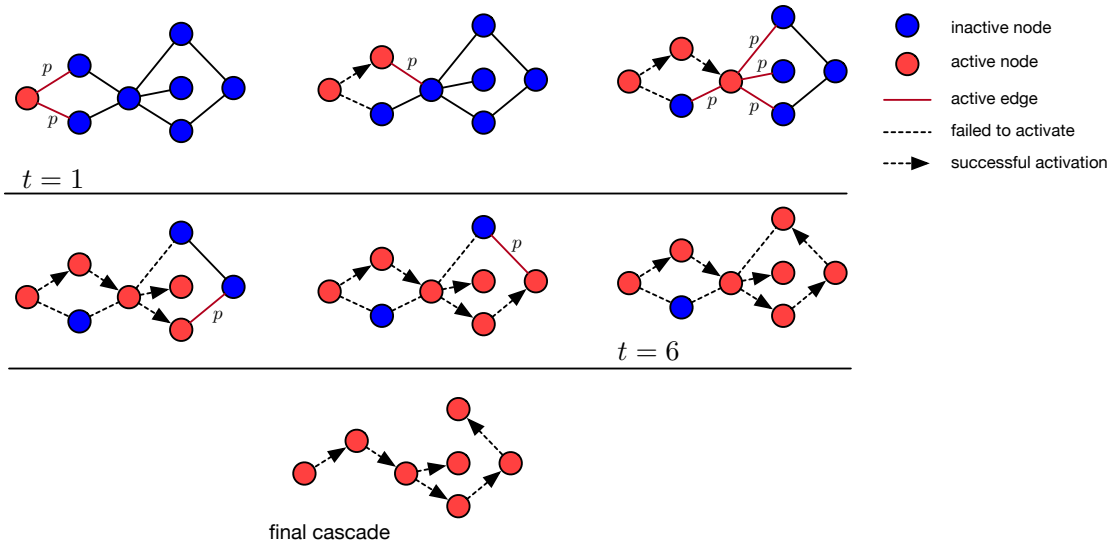


Figure 4: An example of the independent cascade model running on an undirected network of $n = 8$ nodes, where at $t = 1$ a single node is active. At each time step, if a new node u becomes active, then each of its edges (u, v) becomes "active" and with probability p, v becomes active in the next time step; otherwise, that edge becomes inactive forever. The cascade stops when no new nodes become active. The final cascade can then be extracted as the set of edges along which activations passed successfully.

[10] This version clearly described in Kempe, Kleinberg, & Tardos, "Maximizing the spread of influence through a social network." *Proc. 9th KDD*, 137–146 (2003).

[11] Often, we choose these "seed" nodes uniformly at random, but there is a rich literature on selecting the initially active nodes in different ways so as to, e.g., maximize the downstream size of the cascade ("influence maximization") or minimize the probability of not being activated ("information access gap minimization").

This model is often used when thinking about information cascades in social media, e.g., retweet or reshares, or the spread of gossip, where the cost of transmission is low and peer effects are modest.

Linear Thresholds. The LT model has a long history, stretching back at least as far as Granovetter and Schelling,[12] and is one way to formalize the idea of peer pressure, i.e., that a certain fraction of a person's friends need to adopt a behavior before that person will follow suit.

As in the IC model, we again have a directed or undirected network $G = (V, E)$, and a set of states $\Gamma = \{I, A\}$ (inactive and active). In addition, we assign each node a threshold θ_u. (Often we set θ_u to be a uniformly random variable on the unit interval $\sim U(0, 1]$). This node-specific parameter denotes the fraction of u's neighbors that must become active in order for u to become active. We then initialize some subset of the nodes to be active and time proceeds in discrete steps.[13]

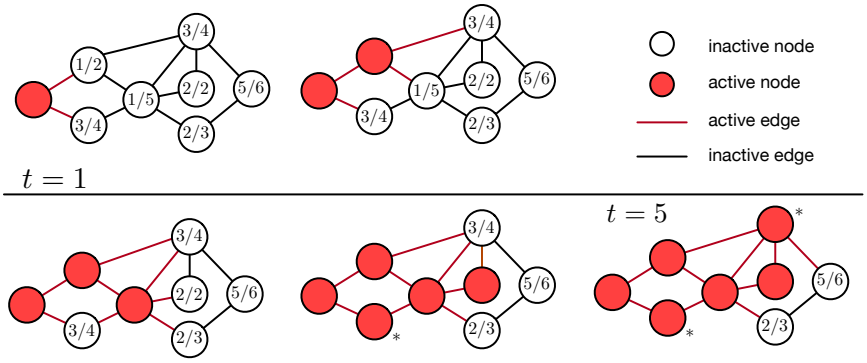


Figure 5: An example of the linear threshold model running on an undirected network of $n = 8$ nodes, with thresholds θ_u marked, and where at $t = 1$ a single node is active. At each time step, if a new node u becomes active, then each of its edges (u, v) becomes "active" and we check each node at the end of every active edge to see if at least a fraction θ_v of its edges are active; if so, v becomes active in the next time step. The cascade stops when no new nodes become active. Because multiple neighbors can be required to activate, there is no final "cascade" in the sense of the IC model. Nodes marked with $*$ became active at a time step well after their first neighbor became active.

In each step, we activate any node u for which the total weight of its active neighbors exceeds its threshold θ_u, i.e., the update rule for the LT model is

$$\sum_{(v,u)\in E} \mathrm{I}[x_v = A] \times 1/k_u \geq \theta_u \ ,$$

[12] See Granovetter. "Threshold models of collective behavior." *Am. J. Sociology* **83**, 1420–1443 (1978), and Schelling, *Micromotives and Macrobehavior*, Norton (1978).

[13] The same ideas as in the IC model for strategically selecting this "seed" set apply here, but the results will differ because the model of spreading is different.

where $\mathrm{I}[x_u = A]$ is an indicator variable for whether the node u is active or not. Time then proceeds until no new activations occur. In this framework, activation occurs when the fraction of neighbors that are activated exceeds a threshold, but we could just as easily make that a weighted fraction, e.g., if some connections to u are more important than others.

Simple vs. complex contagions. Given either model of social spreading, we can measure various things about the "cascade" of activations it produces: what shape does the activation tree have? how deep (longest chain) is the tree? how large (number of activated nodes) is the tree? do high-degree nodes tend to activate early or late in a cascade? etc. We can answer these questions via simulation, given some assumptions for setting the parameters and a choice of network, which allows us to compare and contrast these different models and gain some intuition about how information might spread in reality.

Empirical work studying real information cascades on real social networks suggests that different kinds of social spreading are better modeled by the IC versus the LT model,[14] and that the difference may stem from the cost imposed on the adoption. If the cost is low, e.g., in sharing information, then the IC model appears to be better. But, if the cost is high, e.g., in adopting a behavior like spending money to make a purchase or spending time to carry out an action, then the LT model appears to be better.

In the latter case, the literature sometimes calls these things *complex contagions*, because it can take multiple "attempts" at influence to activate a node, i.e., more than one active neighbor is necessary. Complex contagions (LT models) spread more efficiently through locally dense regions of a network, where there are many short paths between each nearby pairs of nodes, which allows the social behavior to reinforce its influence and exceed a node's threshold. As a result, complex contagions should tend to activate high-degree nodes later in the diffusion.

In the former case, they are sometimes called *simple contagions*, because activation can be done by a single neighbor. Simple contagions (IC models) spread more efficiently through sparsely connected, or locally tree-like regions, and will tend to find the high-degree nodes sooner.

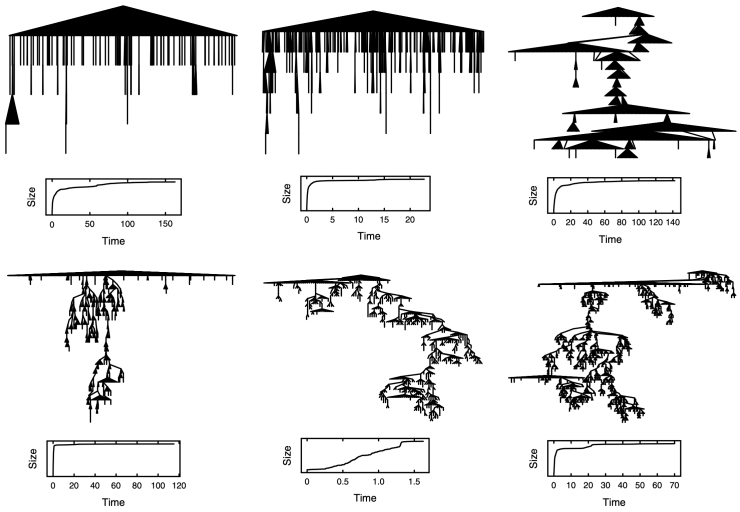
1.3 Cascades on real social networks

There are many interesting questions we can ask about cascades on real social networks. Cascades in epidemiology are considerably harder to study than "information" cascades in online social networks, largely because both activation events and the underlying social network are more readily observable in the latter case than in the former. But, in both settings, we might ask, How predictable is a cascade? which amounts to trying to predict both which individual nodes will be come

[14] See Centola, "The Spread of Behavior in an Online Social Network Experiment." *Science* **329**, 1194 (2010).

activated (infected) next and how large will the cascade eventually be? The particular answer is likely to differ depending on the details of the underlying transmission mechanism, and some evidence suggests that epidemiological cascades are easier to predict than information cascades, in part because of the role that physical, environmental, and structural factors play in epidemics, which are often easier to measure than the social factors that shape social influence.

Twitter is a popular platform for studying information cascades because of its scale, heterogeneity, and availability of data. Writing in 2015, Goel et al. studied roughly 10^9 cascades on Twitter of uniquely identifiable URLs over a 12 month period starting in July 2011. As one might expect for a complex social system, they first find that Twitter information cascades exhibit a very heavy-tailed distribution (possibly power-law shaped) in their final size (number of activated nodes)—the number of reshares or retweets of an original share. The distribution is so skewed that the average number or reshares is just 1.3, and 99% of cascades reach no further than the followers of the seed node. Only a tiny faction—0.025%—of cascades contain 100 or more nodes. (What does this say about the transmission dynamics of information in this setting? Why do you think that is so? Is this pattern more consistent with an IC model or a LT model? How might you decide?)



Notes. For ease of visualization, cascades were restricted to having between 100 and 1,000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade, with time indicated in hours. For visual clarity, the adoption curves terminate at 99% of the final cascade size.

However, if an information cascade does spread, it can grow very large. The figure above (reproduced from Goel et al.) visualizes six large cascades that span a wide range of structural characteristics. The first two illustrate "broadcast" type patterns, with one extremely high-degree node at the root with relatively shallow cascades underneath. The third on the top row shows a sequence of more moderate-sized trees, while the bottom row shows cascades with a number of long "chains" between the bottom and top of the cascade.

Supplemental readings

1. Watts et al., "Multiscale, resurgent epidemics in a hierarchical metapopulation model." *Proc. Natl. Acad. Sci. USA* **102**, 11157 (2005)
https://www.pnas.org/content/102/32/11157

2. Brockmann and Helbing, "The Hidden Geometry of Complex, Network-Driven Contagion Phenomena." *Science* **342**, 1337 (2013)
https://www.science.org/doi/10.1126/science.1245200

3. Monod et al., "Age groups that sustain resurging COVID-19 epidemics in the United States." *Science* **371**, eabe8372 (2021)
https://www.science.org/doi/10.1126/science.abe8372

4. Centola, "The Spread of Behavior in an Online Social Network Experiment." *Science* **329**, 1194 (2010).
https://www.science.org/doi/10.1126/science.1185231

5. Goel et al., "The Structure of Virality of Online Diffusion." *Management Science* **62**, 180–196 (2015).
http://dx.doi.org/10.1287/mnsc.2015.2158