Weighted networks: Applications from power grid construction to crowd control

A Dissertation Presented

by

Thomas Charles McAndrew

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

January, 2017

Defense Date: October 7th, 2016
Dissertation Examination Committee:

James P. Bagrow, Ph.D., Advisor
Chris Danforth, Ph.D., Advisor
Matthew Price, Ph.D, Chairperson
Joshua Bongard, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

Since their discovery in the 1950's by Erdős and Rényi, network theory (the study of objects and their associations) has blossomed into a full-fledged branch of mathematics. Due to the network's flexibility, diverse scientific problems can be reformulated as networks and studied using a common set of tools. I define a network $G = (V, E)$ composed of two parts: (i) the set of objects V, called nodes, and (ii) set of relationships (associations) E, called links, that connect objects in V. We can extend the classic network of nodes and links by describing the intensity of these associations with weights. More formally, weighted networks augment the classic network with a function $f(e)$ from links to the real line, uncovering powerful ways to model real-world applications.

This thesis studies new ways to construct robust micro powergrids, mine people's perceptions of causality on a social network, and proposes a new way to analyze crowdsourcing all in the context of the weighted network model.

The current state of Earth's ecosystem and intensifying climate calls on scientists to find new ways to harvest clean affordable energy. A microgrid, or neighborhood-scale powergrid built using renewable energy sources attached to personal homes, suggest one way to ameliorate this energy crisis. We can study the stability (robustness) of such a small-scale system with weighted networks. A novel use of weighted networks and percolation theory guides the safe and efficient construction of power lines (links, E) connecting a small set of houses (nodes, V) to one another and weights each power line by the distance between houses. This new look at the robustness of microgrid structures calls into question the efficacy of the traditional utility. The next study uses the twitter social network to compare and contrast causal language from everyday conversation. Collecting a set of 1 million tweets, we find a set of words (unigrams), parts of speech, named entities, and sentiment signal the use of informal causal language. Breaking a problem difficult for a computer to solve into many parts and distributing these tasks to a group of humans to solve is called Crowdsourcing. My final project asks volunteers to 'reply' to questions asked of them and 'supply' novel questions for others to answer. I model this 'reply and supply' framework as a dynamic weighted network, proposing new theories about this network's behavior and how to steer it toward worthy goals.

This thesis demonstrates novel uses of, enhances the current scientific literature on, and presents novel methodology for, weighted networks.

Citations

Material from this dissertation was published in the following forms:

McAndrew, Thomas C., Christopher M. Danforth, and James P. Bagrow.. "Robustness of spatial micronetworks." Physical Review E 91.4 (2015): 042813.

and

McAndrew, T. C., Bongard, J. C., Danforth, C. M., Dodds, P. S., Hines, P. D., & Bagrow, J. P.. (2016). “What we write about when we write about causality: Features of causal statements across large-scale social discourse.” ASONAM 16 (2016).

McAndrew, T. C. & Bagrow, J. P.. (2016). “Efficient crowdsource exploration for growing question networks” WSDM 17 (2017) (Submitted).

Dedication

To my wife and dogs, without you I could never reach this high.

Acknowledgments

Thank you to:

- Mum, Pop, and sister: For your constant love and support.
- Andi Elledge: For great conversations, and helping organize my life.
- Leonardo's and Junior's Pizza: SCRAPS would never meet without your pizza.
- Mark Wagy: I would not progress this fast without our enlightening conversations.
- Emily Cody: For excellent scientific thoughts, after pizza.
- Andy Reagan: Our dog-coding sessions were always helpful.
- Nathan Lemons and Aric Hagberg: You both made me a better scientist.
- Jake Williams: Alot of mathematics in this dissertation was sketched on your whiteboard.
- Dilan Kiley: For being a good friend.
- Peter Dodds: Friend, Mentor, and Complexity enthusiast
- Joshua Bongard: Excellent conversations about artificial intelligence
- Yves Dubief: Excellent conversation on all things, especially baked goods.
- Chris Danforth: For inspiring my journey to UVM
- James Bagrow: For molding me into a better scientist. I know this wasn't easy.

List of Tables

List of Figures

Table of Contents

Chapter 1

Introduction

From discovering infrastructure weaknesses in our internet [6, 29, 45] to explaining the intensity of our social connections [66, 16, 36, 117, 167], complex networks permeate many branches of science [144, 76, 8, 61, 166]. Fields like power systems, social network analysis, and neuroscience easily lend themselves to network study achieving important breakthroughs, while many other fields adopted network theory as a way to mold onerous scientific problems into solvable forms. It is the flexibility of network analysis, a simple description of objects and associations, that drives a plethora of new applications, and with numerous applications and benefits to future scientific endeavor, studying and contributing to the theory and applications of networks becomes a worthy goal of this thesis.

1.1 Network Theory

From the 1700's to just before Erdős and Rényi's publication in the 1960's and motivated by practical problems, a network was considered a set of nodes (V) bound together by a set of static links (E). Euler was the first to study networks with his 'Bridge of Konigsburg' problem [52] which asked if a single person could cross every bridge in Konigsburg exactly once and return to their starting position. Further applications, like the enumeration of chemical polymers and the study of voltage across an electrical circuit motivated great minds such as Kirchoff, Cayley, and Polya. While these types of applications were suitable for static networks, Erdős and Rényi's innovation [48, 47, 49] defined a network as a set of objects called nodes accompanied by a set of **probabilistic** relationships between nodes called links. This monumental mathematical discovery, coined the Erdős Rényi (ER) network $G(N, p)$, defines a fixed set of N nodes and a uniform probability p that any two nodes can connect.

One important finding suggests a network's degree distribution informs us about this network's function. A network's degree distribution assigns a probability $p(k)$ to each possible degree (number of links) a node can take, characterizing the structure of the network. Two important degree distributions discussed in this thesis are the powerlaw and Erdős-Rényi distribution. When a network follows a powerlaw degree distribution

$$(p(k|\gamma) \sim k^{-\gamma}) \tag{1.1}$$

very few (many) nodes contain a sizable (modest) number of links, and this imbalance implies the overall connectivity of the network relies on a small number of important nodes. This distribution was introduced by Yule, Polya, and further generalized by Simon [154] who showed the 'rich-get-richer' story leads to a powerlaw distribution. Simon's dynamical model that gives rise to a powerlaw distribution is simple. Begin with a single object of a specific genre. At each time step t, a new object enters the system and takes on a never before seen genre with probability ρ, or a previous genre with probability $1 - \rho$. If the object must take a previous genre, it does so with probability proportional to the size of objects already in each genre (i.e. a genre with double the number of members has double the probability to accrue new objects of the same genre). This simple mechanism was later re-discovered by Barabasi and Albert in the context of networks and their degree distribution.

Contrary to a power law degree distribution, an Erdős-Rényi network's degree distribution has a broader Poisson distribution

$$(p(k|\lambda) \sim \frac{\lambda^{-k}}{k!}). \tag{1.2}$$

This distribution arises from taking a fixed number of nodes and adding in a fixed number of links between nodes at random. As one holds the proportion of links to nodes constant and increases the network size, a Poisson distribution arises.

This new field of stochastic networks opened up new doors to applications like the break down of chemical bonds [125, 107] and the connectedness of social networks [57, 104, 124, 64], but these networks lack further information about the intensity of relations between nodes and links.

1.2 Weighted Networks

Weighted networks extend probabilistic networks by associating strengths to the existing relationships between nodes. With weighted links, the structure of the links and their intensity become important in describing the system. More formally, we define a weighting function, $f(i, j)$, that takes as input a link between any two objects i and j, and represents their strength of association with a real value. Past work attempts to discover how the connective structure of links and the heterogeneity of weights on these links implies functional changes to a system [17, 158, 41].

Arguably the first work on weighted networks begins with Granovetter's strength of weak ties [66]. Granovetter supposed the relationship strength between two individuals consists of a linear combination of length of time spent together, emotional intensity, and intimacy, and further hypothesized that the stronger the bond between two people the larger the proportion of common friends they share. Granovetter hypothesized our weaker links are the key to extending our social network. These people on the periphery of our social network provide us with diverse information, and new contact compared to our strongly bonded homogenuous group of friends. Since Granovetter's work in social phenomena, we find numerous past applications of weighted networks illustrated with applications such as: infrastructure and construction, genetics and the brain, epidemiological spread of disease, and social influence between people.

Most civics research uses points of interest in the system as nodes and links two points when people commute between them. Extending this model to weighted networks, past authors commonly weight links by the volume of (or capacity to handle) traffic between the two corresponding points [19, 68, 60, 174], and find connections between the linking pattern of networks and weights. For instance, Xu et.al [174] find airports with the most traffic connect to similar high-traffic hubs, but also many low traffic smaller cities. Genetics use weighted networks as models of gene interactions in the body [76, 151]. Most often, authors define genes as nodes, associations between genes as links, and weight pathways on the strength of activation (positive weight) or inactivation (negative weight). Similar in structure, neuroscience defines weights between regions of the brain based on blood flow patterns witnessed with functional magnetic resonance imaging (fMRI), andrelate disease with particular network metrics. Epidemiologists have modeled the spread of disease using weighted networks, removing links with probability based on the strength of connection between contacts, and measuring the final size of the number of infected as the average number of nodes in the largest connected component [115, 89, 153]. Sociologists also study weighted networks, often in the form of personal bonds between a social network of individuals.

Although civic studies, biology, and social science have a long history of publishing on weighted networks, crowdsourcing literature lacks weighted network models, and instead rely on Markov decision processes to form applied and theoretical results. Markov decision processes start with a set of States S the system takes, set of actions A a user can take to change the system state, the probability $P(s|a)$ of moving to state a given the system is in state s, and the reward $R(s, a)$ for taking action a when in state s. This model provides a solid mathematical framework for scientists to prove optimality conditions for algorithms in specific state systems, but a weighted network approach may open new doors.

This thesis work builds on weighted network methodology and its applications to better model the structure of small power grids, discover linguistic markers of causal language, and control a crowd of workers in a dynamic evolving network of tasks.

1.3 Structural Vulnerability and Robustness

Cascading failure and infrastructure vulnerability are important topics studied by physicists and engineers [42, 33, 9, 173], many using weighted networks to describe the system of loads and connections. Two major discoveries show opposing views on how we should construct power grids; the first [4] links the degree distribution of the power grid (structure) to robustness (function), and illustrates robust grids distribute power lines more evenly than vulnerable grids. The second [28] discovery shows negative aspects of building power grids with broad degree distributions when they are interdependent on other external infrastructure.

In the case of a powergrid, the degree distribution relates the possible number of load-to-load connections. A flatter degree distribution in a powergrid spreads responsibility of the entire network over a larger set of generators and strengthens overall connectivity. It was shown [4] a power grid constructed with a (skewed) powerlaw degree distribution becomes vulnerable under targeted attack, but stays robust to random attack. On the other hand, a broader degree distribution breaks apart more easily under random attack, while targeted attacks do not pose a risk.

A targeted attack orders nodes from highest degree to lowest, and destroys nodes one at a time descending down the list. This causes powerlaw distributed networks to lose their most important nodes (highest degree) first, quickly fragmenting large portions of the network. By equalizing the number of links each node contains, broader distributed networks resist targeted attack due to the difference from one node's degree to the next is not as drastic as a powerlaw distributed network. In random attack, powerlaw distributed networks have very few nodes with many connections, and the probability of hitting these valuable nodes is much smaller than attacking a more vital node in a broadly distributed network. This push and pull of structure, coupled with attack strategy, assumed the power grid an independent entity and non-reliant on other infrastructure.

This work was extended [28] by considering two interdependent networks, for example the power grid and gas network, that rely on one another to function. The model considered a within-network degree distribution ($p()$) and between-degree distribution (). The destruction model chooses a node from network A. In the first step, the authors remove this node and neighboring within-network links. Any nodes in network B linked to the attacked node in A are removed, and any links that connect disparate sections of network A are removed in network B, effecting network

B's infrastructure. The results of interdependent structure's impact on cascading failure conflict with previous work on isolated power grid structure; broader degree distributions between networks more heavily depend on one another, and losing a link from one network percolates through both networks. Skewed distributions in both networks depend less on another, and although a crucial node in one network (A) is destroyed, the lack of reliance on the separate network (B) prevents a cascading failure.

Gastner and Newman consider weighted networks while modeling construction [58, 60]. In their work, they consider constructing a network by placing links one by one, minimizing average pairwise distance between any two nodes while also staying under budget. Two weights are appended to each link between nodes i and j, the euclidean distance (d_{ij}) used to compute the budget as

$$\text{Budget} = \sum_{(ij)\in E} d_{ij} \tag{1.3}$$

where E represents the set of all links, and an effective distance

$$\lambda\sqrt{n}d_{ij} + (1-\lambda), \tag{1.4}$$

with $\lambda \in [0,1]$ tuning construction of a topological network (small λ), or geometric graph (large λ). The authors of [58] recover similar attribute to real-world topological (network of airports) and planar graphs (road network). Authors of all three previous works, and many others, show concrete properties of these networks by appealing to the limit as the network grows to infinity. Our work considers the difficulties small networks pose on construction, and we contribute to the conversation of power grid structure and vulnerability by modeling microgrids. A microgrid consists of a small array of houses connected together by a common source of energy [99, 7, 109]. The microgrid idea, while used in military and industrial settings, has entered the limelight as a sustainable alternative to the traditional hierarchical structure of the power grid.

Our work builds on Gastner and Newman's network construction [58] by studying how a small network's construction can influence robustness. Specifically, we build networks the same as in [58], but now percolate the network by moving to each link, destroying this link with probability

$$p_{ij} \sim q d_{ij}^{\alpha} \tag{1.5}$$

where $\alpha \in [1,\infty)$ and $q \in [0,1]$ serve as 'damage' parameters, and measure the largest connected component that remains. Our main finding states that building non-spatial networks (small λ) forms long links, and when these links are destroyed, quickly fragments the network. On the other hand, networks with a more even distribution of links maintain a higher robustness.

1.4 Investigating Causality on Social networks

Social networks, with people represented as nodes and links demonstrating a social connection between two people, have increasing importance in many scientific disciplines. Understanding how people exchange goods and services motivates economists to explore social networks, while advertising attempts to abuse our social connections for higher profits margins. The internet allows sociologists, like researchers at Facebook [46], a limitless amount of data to probe complicated social hypothesis. Better understanding social phenomena may better predict terrorist activity [100, 53], understand how we best motivate people toward using energy responsibly [71, 114, 132], or even identifying novel causal links [67, 145].

Causality, the holy grail of science, remains hotly debated among scientists and philosophers alike; previous linguistic work in this area focuses on (i) automatically discovering new causal words and grammatical structure, (ii) mining causal links from common textual sources, and (iii) predicting future events from a data-driven causal model.

Most work on discovering linguistic patterns in text start with a template pattern [63, 129, 62]. The causal templates patterns include sentences surrounded by the word cause or any of its conjugations. Finding similar causal verbs amounts to finding statistically unlikely similarities between other verb's placement in sentences [62]. Many of the semantic features of text used for causal verb finding, like part-of-speech tagging, rely on machine learning algorithms.

A similar view on finding causal verbs was applied to discovering causal links between two concepts. Previous authors first determine whether or not a sentence is causal and then locate the two noun phrases on each side of the verb. Statistically frequent noun-phrase pairs that occur on either side of the verb are described as potential causal links [63]. Other techniques focus more intensely on the temporal axiom, and one project in particular relates time-stamped textual topics to non-textual time-stamped time series [93]. The authors from [93] use Latent Dirichlet Allocation (LDA) to uncover topics within time stamped text. Causal relations are supposed by correlating lagged topic coverage's with non-textual time series, for example stock prices or the 'winner-takes-all' Iowa market.

We build on previous causal work in computational linguistics by studying people's informal perceptions of causality. Our project combines frequency statistics on linguistic components, document level topic analysis, and corpus wide sentiment analysis to tease out differences between causal language in a social network compared to random text.

1.5 Crowdsourcing

The end of this thesis focuses on the weighted networks role in crowdsourcing. The Crowdsourcing paradigm develops a solution by breaking the problem into manageable pieces a human can complete and dealing these pieces out in parallel. A researcher benefits from crowdsourcing a problem, over a more computational solution, in two ways: (i) humans can solve these tasks much easier and more accurately than a computer and (ii) the task may require human ingenuity. the reCaptcha system and galaxy zoo are good examples of a human's ability to solve a problem a computer cannot. The reCaptcha algorithm was produced after optical character recognition (OCR) systems left specific words unidentified in a corpus of text. These missing words are identified by giving them to individuals over the internet and asking them to identify the word. After receiving three consecutive and agreeing answers the word is considered identified. After a brief tutorial, Galaxy zoo asks users to classify different types of pictures taken from space into one of many types of galaxies. Millions of classifications are made by 'citizen-scientists' per hour, so much so, most astronomers have trouble keeping up with the data. Human's are excellent at visual classification and inspection, making projects like reCaptcha and galaxy zoo instantly successful. Other work in crowdsourcing focuses on our creativity capacity, rather than our ability to classify images visually. Threadless and Dell's product storm showcase two Crowdsourcing platforms that harness human ingenuity. An online retail company, Threadless allows customers to design and purchase their custom made new clothing products. In the same vein, IdeaStorm helps the computer company Dell better design computers for their consumers by letting them post creative improvements to Dell, and vote on others ideas as good or bad. Dell can then find the most voted on improvements to computer and consider altering their production line.

Past literature focuses on using humans to complete tasks accurately and efficiently (minimal expenditure of resources), and theoretical work couches itself in the Markov Decision process. The traditional crowdsourcing Markov decision process (MDP) links the state of the system with the behavior of the workers. The actions that the task-requester can make are monetary payments toward workers, and are often based on heuristic criterion. Finally, most formulations learn the state transitions (worker behavior) from empirically collected data.

Crowdsourcing MDP literature designs optimal policies meant to pay workers to acquire meaningful answers, but not overspend. One example measures a meaningful answer (and worker) with Information Gain, computed as the entropy (H) before the answer minus the current information. A positive (negative) number represents a gain (loss) in task certainty. Payment schedules vary among studies but usually include past worker answers, certainty of answers to tasks, and the probability of transitioning to specific system states (either positive of negative). [82, 35, 79]

Instead of taking the traditional Markov Decision Approach, my work models crowdsourcing as a weighted network where links represent tasks for workers to complete, and nodes symbolize pieces of a task that can may be used in other future tasks. Weights for a task (link) account for worker's answers, for example if the task is a 'Yes or No' question the number of 'Yes's and 'No's will weight the link. In a novel way, I map the crowdsourcing paradigm onto a weighted network, and can use weighted network machinery to answer questions about the overall system of workers and task completion.

A second contribution to crowdsourcing science was coupling human ingenuity with accuracy and efficiency. Most crowdsourcing models are static in the number of tasks to complete. Instead of assuming all tasks for a problem have been specified, I assume hidden tasks for discovery and allow the workers themselves to supply guesses at these hidden tasks. The system's task-set now evolves. The crowdsourcer must now balance exploration of the network through worker proposals, with using more worker answers on tasks that are undecided.

This thesis ties together disparate applied fields with weighted network theory. Percolation theory and Monte Carlo simulation show we can build robust networks when constrained under a budget. Extracting information from the twitter social network we find marked differences in how people use causal language. Finally, weighted networks offer a natural way to describe and theorize about a team of workers working together on a complex, growing set of questions. This work draws the starting-line from which other scientists may run.

1.6 Summary

In Chapter 2 this thesis demonstrates, in detail, how weighted networks can help determine a fully connected microgrid that remains as robust as possible under random failure, and rediscovers classical results in grid vulnerability by random attack on links proportional their length (weight). We randomly place 50 nodes in the unit square and build connected networks that must (i) stay under budget and (ii) remain fully connected. Robustness of each grid is examined by weighting the failure of links proportional to their euclidean distance. We go on to show nodes that contain more links are more likely to contain longer links, becoming key points of attack, and offer advice to practitioners and future builders of microgrids.

Chapter 3 examines causality on the twitter social network and finds key differences in how everyday people discuss causality compared to other subjects. Collecting approximately 1 million causal tweets and time-matched random tweets, we find differences in how people use causal unigrams, parts of speech, and named entities. We continue to contrast causal with control tweets and group causal tweets into distinct topics including: news, medicine, and social phenomena. Finally, we analyze the sentiment of causal versus control tweets and observe when people write about causality they write more negatively.

Chapter 4 uses weighted networks to propose a novel "reply-and-supply" aspect to crowd sourcing, and suggests probability matching to leverage exploring, through human innovation, untouched parts of a task while at the same time exploiting the crowd to gain more confidence on tasks that are less sure. Our crowdsourcing example starts from a single word pair (words i and j) linked together, and asks workers "are these synonyms" . After the workers answers yes or no, they are allowed to contribute and explore the problem by suggesting a potential synonym to word i, synonym to word j or a common synonym to both words i and j. We weight each link in the network by the distribution of yes/no answers, and pick the next link to give a worker based on link uncertainty.

In each chapter we show a different aspect of how weighted networks effectively model and solve complex applied problems. The conclusion ties each chapter together into a cohenerent discourse on weighted network theory, social phenomena, and crowdsourcing. This thesis inspires robustness micro grid construction, provides insight into human perception of causality, and develops a new theory on crowdsourcing through the lens of weighted networks.

Chapter 2

Building robust infrastructure via weighted networks

Abstract

Power lines, roadways, pipelines and other physical infrastructure are critical to modern society. These structures may be viewed as spatial networks where geographic distances play a role in the functionality and construction cost of links. Traditionally, studies of network robustness have primarily considered the connectedness of large, random networks. Yet for spatial infrastructure physical distances must also play a role in network robustness. Understanding the robustness of small spatial networks is particularly important with the increasing interest in microgrids, small-area distributed power grids that are well suited to using renewable energy resources. We study the random failures of links in small networks where functionality depends on both spatial distance and topological connectedness. By introducing a percolation model where the failure of each link is proportional to its spatial length, we find that, when failures depend on spatial distances, networks are more fragile than expected. Accounting for spatial effects in both construction and robustness is important for designing efficient microgrids and other network infrastructure.

2.1 Introduction

The field of complex networks has grown in recent years with applications across many scientific and engineering disciplines [5, 120, 121]. Network science has generally focused on how topological characteristics of a network affect its structure or performance [5, 29, 120, 159, 40, 121]. Unlike purely topological networks, spatial networks [18] like roadways, pipelines, and the power grid must take physical distance into consideration. Topology offers indicators of the network state, but ignoring the spatial component may neglect a large part of how the network functions [168, 13, 28, 4]. For spatial networks in particular, links of different lengths may have different costs affecting their navigability [96, 139, 30, 32, 31, 102] and construction [142, 58, 59, 60, 162].

Percolation [157] provides a theoretical framework to study how robust networks are to failure [29, 123, 11, 150, 147]. In traditional bond percolation, each link in the network is independently removed with a constant probability, and it is asked whether or not the network became disconnected. Theoretical studies of percolation generally assume very large networks that are locally treelike, often requiring millions of nodes before finite-size effects are negligible. Yet many physical networks are far from this size; even large power grids may contain only a few thousand elements.

There is a need to study the robustness of small spatial networks. Microgrids [99, 155, 69, 87] are one example. Microgrids are small-area (30–50 km), standalone power grids that have been proposed as a new model for towns and residential neighborhoods in light of the increased penetration of renewable energy sources. Creating small robust networks that are cost-effective will enable easier introduction of the microgrid philosophy to the residential community. Due to their much smaller geographic extent, an entire microgrid can be severely affected by a single powerful storm, such as a blizzard or hurricane, something that is unlikely to happen to a single, continent-wide power grid. Thus building on previous work, we consider how robustness will be affected by spatial and financial constraints. The goal is to create model networks that are both cost-effective, small in size, and at the same time to understand how robust these small networks are to failures.

The rest of this paper is organized as follows. In Sec. 2.2 a previous model of spatial networks is summarized. Section 2.3 contains a brief summary of percolation on networks, and applies these predictions to the spatial networks. In Sec. 2.4 we introduce and study a new model of percolation for spatial networks as an important tool for infrastructure robustness. Section 2.5 contains a discussion of these results and future work.

2.2 Modeling infrastructure networks

In this work we consider a spatial network model introduced by Gastner & Newman [59, 58, 60], summarized as follows. A network consists of $|V| = N$ nodes represented as points distributed uniformly at random within the unit square. Links are then placed between these nodes according to associated construction costs and travel distances. The construction cost is the total Euclidean length of all edges in the network, $\sum_{(i,j)\in E} d_{ij}$, where d_{ij} is the Euclidean distance between nodes i and j and E is the set of undirected links in the network. This sum represents the capital outlay required to build and maintain the network. When building the network, the construction cost must be under a specified budget. Meanwhile, the travel distance encapsulates how easy it is on average to navigate the network and serves as an idealized proxy for the functionality of the network. The degree to which spatial distance influences this functionality is tuned by a parameter λ via an "effective" distance

$$d_{\text{eff}}(i,j) = \sqrt{N}\lambda d_{ij} + (1-\lambda).$$

Tuning λ toward 1 represents networks where the cost of moving along a link is strongly **spatial** (for example, a road network) while choosing λ closer to 0 leads to more **non-spatial** networks (for example, air transportation where the convenience of traveling a route depends more on the number of hops or legs than on the total spatial distance). To illustrate the effect of λ, we draw two example networks in Fig. 2.1. Finally, the travel distance is defined as the mean shortest effective path length between all pairs of nodes in the network. Taken together, we seek to build networks that minimize travel distance while remaining under a fixed construction budget, i.e. given fixed node positions, links are added according to the constrained optimization problem

$$\begin{aligned} \min \frac{1}{\binom{N}{2}} \sum_{s,t\in V} \sum_{(u,v)\in\Pi(s,t)} d_{\text{eff}}(u,v) \\ \text{subject to} \sum_{(i,j)\in E} d_{ij} \leq \text{Budget}, \end{aligned} \tag{2.1}$$

where $\Pi(s,t)$ is the set of links in the shortest effective path between nodes s and t, according to the effective distances d_{eff}. (The factor $\binom{N}{2}^{-1}$ does not affect the optimization.) This optimization was solved using simulated annealing (see App. 2.6.1 for details) with a budget of 10 (as in [58]) and a size of $N = 50$ nodes. We focus on such a small number of nodes to better mimic realistic microgrid scales. In this work, to average results, 100 individual network realizations were constructed for each λ.

An important quantity to understand in these networks is the distribution of Euclidean link lengths. If edges were placed randomly between pairs of nodes, the

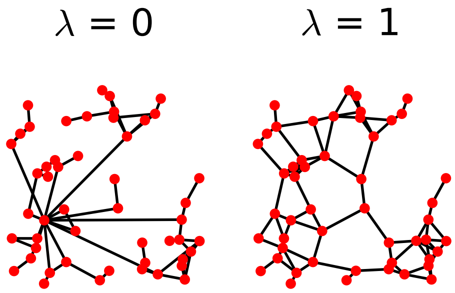


Figure 2.1: (Color online) Two optimized spatial networks with the same node coordinates illustrate how λ influences network topology. The ***non-spatial*** *case $\lambda = 0$ shows long-range hubs due to the lack of restriction on edge distance; the* ***spatial*** *case $\lambda = 1$ lacks expensive long distance links leading to a more geometric graph. As examples, the non-spatial case may correspond to air travel where minimizing the number of flights a traveler takes on a journey is more important than minimizing the total distance flown, while the spatial case may represent a road network where the overall travel distance is more important than the number of roads taken to reach a destination.*

lengths would follow the square line picking distribution with mean distance $\langle d \rangle \approx 0.52141$ [170]. Instead, the optimized network construction makes long links costly and we observe (Fig. 2.2) that the probability distribution $P(d)$ of Euclidean link length d after optimization is well explained by a gamma distribution, meaning the probability that a randomly chosen edge has length d is

$$P(d) = \frac{1}{\Gamma(\kappa)\theta^{\kappa}} d^{\kappa-1} e^{-d/\theta}, \tag{2.2}$$

with shape and scale parameters $\kappa > 0$ and $\theta > 0$, respectively. A gamma distribution is plausible for the distribution of link lengths because it consists of two terms, a power law and an exponential cutoff. This product contains the antagonism between the minimization and the constraint in Eq. (2.1): Since longer links are generally desirable for reducing the travel distance, a power law term with positive exponent is reasonable, while the exponential cutoff captures the need to keep links short to satisfy the construction budget and the fact that these nodes are bounded by the unit square. See Fig. 2.2.

The fit of the gamma distribution was tested statistically using the Kolmogorov-Smirnov test [112]. The test failed to reject the null hypothesis ($p > 0.05$) that the distances follow a gamma distribution in 99.15% of all networks. This provides strong evidence in favor of Eq. (2.2).

The network parameters were chosen under conditions that were general enough to apply to any small network, for instance a microgrid in a small residential neighborhood. The choices of 50 nodes and a budget of 10 were also made in line with previous

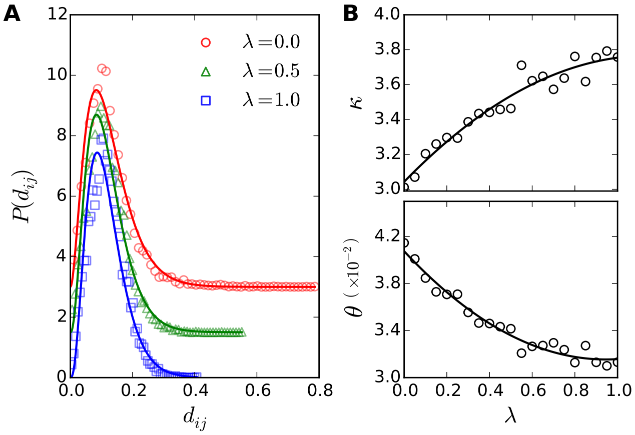


*Figure 2.2: (Color online) The distributions of Euclidean link lengths d_{ij} between nodes i and j are well explained for all λ by gamma distributions, i.e. $P(d_{ij}) \propto d_{ij}^{\kappa-1} e^{-d_{ij}/\theta}$. (**A**) Maximum likelihood estimates of $P(d_{ij})$ for multiple λ. Two distributions are **shifted vertically** for clarity. (**B**) The gamma parameters κ, θ as functions of λ. Quadratic fits provide a guide for the eye.*

studies [58] of this network model to balance small network size with a budget that shows the competition between travel distance minimization and construction cost constraint [59, 58].

2.3 Robustness of physical infrastructure

Percolation theory on networks studies how networks fall apart as they are sampled. For example, in traditional bond percolation each link in the network is independently retained with probability p (equivalently, each link is deleted with probability $q = 1 - p$). This process represents random errors in the network. The percolation threshold q_c is the value of q where the giant component, the connected component containing the majority of nodes, first appears. Infinite systems exhibit a phase transition at q_c, which becomes a critical point [157]. In this work we focus on small **micronetworks**, a regime under-explored in percolation theory and far from the thermodynamic limit invoked by most analyses. In our finite graphs, we estimate q_c as the value of q that corresponds to the largest S_2, where S_n is the fraction of nodes in the n^{th} largest connected component (Fig. 2.3). In finite systems the second largest component peaks at the percolation threshold; for $q > q_c$ the network is highly disconnected and all components are small, while for $q < q_c$ a giant component almost surely encompasses

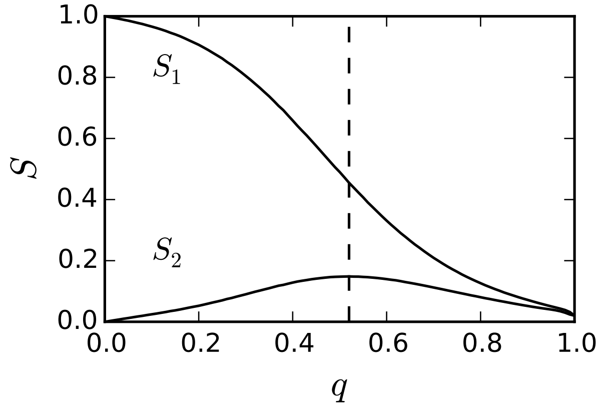


Figure 2.3: The fractions of nodes in the first and second largest components, S_1 and S_2, respectively, as a function of link deletion probability q. In finite systems the percolation threshold q_c can be estimated from the maximum of S_2 (dashed line). This example used optimized networks with $\lambda = 1/2$.

most nodes and S_2 is forced to be small. Note that it is also common to measure the average component size excluding the giant component [74, 157].

For the case of uniformly random link removals (bond percolation) it was shown that the critical point occurs when q is such that $\langle k^2 \rangle / \langle k \rangle = 2$ [37, 119], where $\langle k \rangle$ and $\langle k^2 \rangle$ are the first and second moments of the percolated graph's degree distribution, respectively. We denote this theoretical threshold as $\tilde{q}_\mathrm{c}$ to distinguish this value from the q_c estimated via S_2. Computing this theoretical prediction for the optimized networks (Sec. 2.2) we found $\tilde{q}_\mathrm{c}$ between 0.66 and 0.71 for the full range of λ (Fig. 2.4). In contrast, q_c estimated via S_2 has lower values between 0.48 and 0.54 (Fig. 2.4). It is important to note that the derivation of this condition for $\tilde{q}_\mathrm{c}$ makes two related assumptions that are a poor fit for these optimized spatial networks. First, the theoretical model studies networks whose nodes are connected at random. This assumption does not hold for the constrained optimization (Eq. (2.1)) we study. Second, this calculation neglects loops by assuming the network is very large and at least locally treelike. For the small, optimized networks we build this is certainly not the case. These predictions for the critical point $\tilde{q}_\mathrm{c}$ do provide a useful baseline to compare to the empirical estimates of q_c via S_2.

2.4 Modeling infrastructure robustness

The work by Gastner and Newman [59] showed the importance of incorporating spatial distances into the construction of an infrastructure network model. With

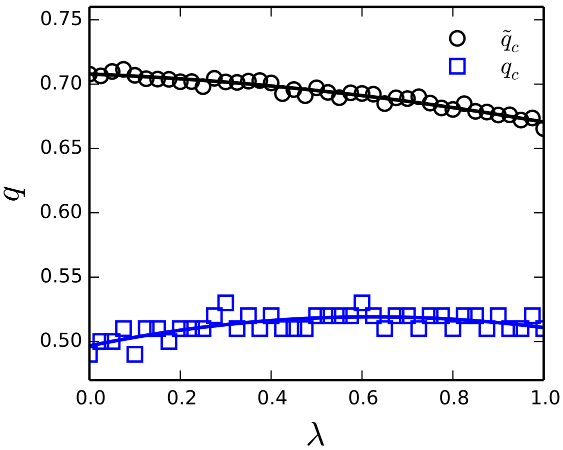


Figure 2.4: (Color Online) For an infinite, uncorrelated network, percolation occurs at the sampling probability for which $\langle k^2 \rangle / \langle k \rangle = 2$ [37]. We computed this predicted critical point $\tilde{q}_{\mathrm{c}}$ for each λ finding $\tilde{q}_{\mathrm{c}}$ between 0.66–0.72. In comparison, for finite networks we used the size of S_2 to estimate q_{c} and found it between 0.48–0.54. Quadratic fits provides a guide for the eye.

physical infrastructure we argue that it is important to also consider spatial distances when estimating how robust a network is to random failures. For example, consider a series of power lines built in a rural area where trees are scattered at random. In a storm trees may fall and damage these lines, and one would expect, all else being equal, that one line built twice as long as another would have twice the chance of a tree falling on it and thus failing.

Motivated by this example, an intuitive model for how links fail would require an increasing chance of failure with length. The simplest model supposes that the failure of a link is directly proportional to length, i.e., that each unit length is equally likely to fail. With this in mind we now introduce the following generalization of bond percolation: Each link (i, j) independently fails with probability $\min(1, Q_{ij})$, where

$$Q_{ij} = qM \frac{d_{ij}^{\alpha}}{\sum_{(i,j)\in E} d_{ij}^{\alpha}} = q \frac{d_{ij}^{\alpha}}{\langle d^{\alpha} \rangle}, \tag{2.3}$$

$q \in [0, 1]$ is a tunable parameter that determines how many edges from 0 to $|E| = M$ will fail on average, and the parameter α controls how distance affects failure probability. We naturally recover traditional bond percolation ($Q_{ij} = q$) when $\alpha = 0$ and $\alpha = 1$ corresponds to the case of constant probability per unit length. See Fig. 2.5 for example networks illustrating how Q_{ij} depends on d_{ij} and α.

Given the gamma distribution of link lengths, the distribution of $y \equiv d^{\alpha}$ is (when

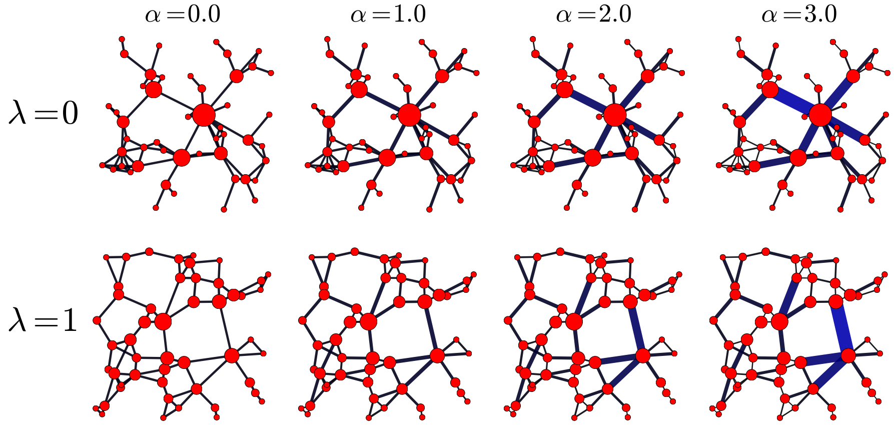


Figure 2.5: (Color online) A non-spatial ($\lambda = 0$) and spatial ($\lambda = 1$) network with multiple values of α showing how to tune the role spatial distance plays in percolation. Here the width and color of a given edge (i, j) are proportional to failure probability $Q_{ij} \sim d_{ij}^{\alpha}$ (2.3) and node size corresponds to the number of effective shortest paths through nodes, with the same scales used across all network diagrams. Increasing α leads to failure probability becoming more concentrated on the links connected to a small number of hubs, with the effected hubs being more central (in terms of shortest paths) for the non-spatial network ($\lambda = 0$) than for the more geometric network.

$\alpha > 0$)

$$P(y) = \frac{1}{\alpha\Gamma(\kappa)\theta^{\kappa}} y^{\kappa/\alpha - 1} \exp\left(-\frac{y^{1/\alpha}}{\theta}\right) \tag{2.4}$$

with mean

$$\langle d^{\alpha} \rangle = \frac{1}{\alpha\Gamma(\kappa)\theta^{\kappa}} \int_0^{\infty} z^{\kappa/\alpha} e^{-z^{\frac{1}{\alpha}}/\theta} \, \mathrm{d}z = \theta^{\alpha} \frac{\Gamma(\kappa + \alpha)}{\Gamma(\kappa)}. \tag{2.5}$$

When $\alpha = 1$, the above distribution (2.4) will reduce to the original distribution $P(d)$, Eq. (2.2).

With the above failure model and the distribution $P(d)$, we may express the probability $P(Q_{ij})$ that a randomly chosen edge (i, j) has failure probability Q_{ij} as

$$P(Q_{ij}) = \frac{1}{\alpha\Gamma(\kappa)\theta^{\kappa}} \left(\frac{\langle d^{\alpha} \rangle}{q}\right)^{\kappa/\alpha} Q_{ij}^{\frac{\kappa}{\alpha} - 1} \exp\left[-\frac{1}{\theta}\left(\frac{\langle d^{\alpha} \rangle}{q} Q_{ij}\right)^{\frac{1}{\alpha}}\right]. \tag{2.6}$$

This distribution has mean $\langle Q_{ij} \rangle = q$. (However, the true mean failure probability is $\langle \min(1, Q) \rangle \leq \langle Q \rangle$ which leads to a small correction, easily computed, as q gets closer to 1.) Note that, while the mean does not rely on the distances of edges, α (and $\langle d^{\alpha} \rangle$) do play a role in higher moments. For example, the variance of Q is $\sigma^2(Q) = q^2 \left(\mathrm{B}(\alpha, \kappa)/\mathrm{B}(\alpha, \alpha + \kappa) - 1\right)$, where $\mathrm{B}(x, y)$ is the Beta function.

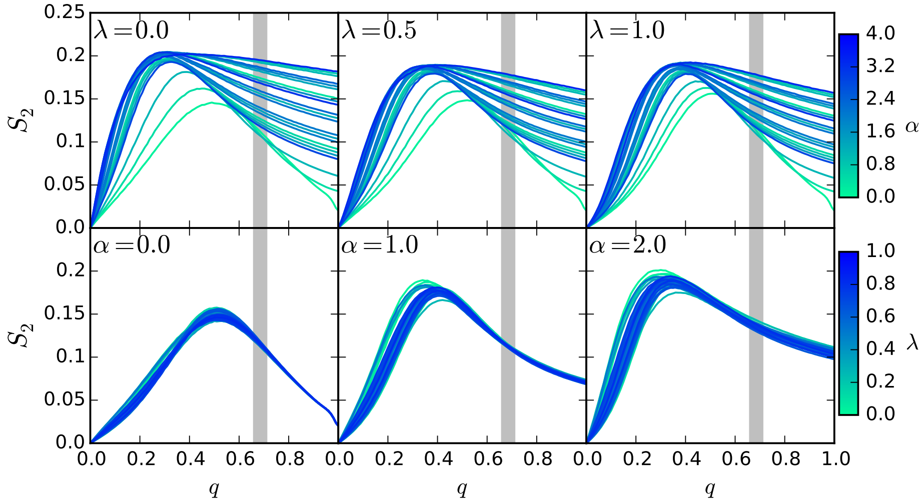


*Figure 2.6: (Color online) In each panel, S_2 (fraction of nodes in the 2nd largest component) curves are shown with the range of the theoretical threshold $\tilde{q}_c$ shown in gray. Higher values of α make failures depend more strongly on distance, while changing λ adjusts a network's form from non-spatial (small λ) to spatial (large λ). (**Top row**) Regardless of λ, larger values of α tend to shift the peak of S_2 towards lower q, leading to less robust networks. (**Bottom row**) Different values of λ for a given α lead to shifted S_2 profiles, but the shift is less prominent. Regardless of the parameters, spatial networks are more fragile than predicted from theory [37, 38]. While both parameters influence the robustness of these spatial networks, α plays a stronger role than λ. Note that with our definition of Q (2.3), S_2 may remain finite as $q \to 1$.*

To study this robustness model we percolate the infrastructure networks by stochastically removing links (i, j) with probabilities Q_{ij} (Eq. (2.3)) for $0 < q < 1$ and $0 < \alpha < 4$. In Fig. 2.6 we plot S_2 vs. q for various combinations of α and λ. Importantly, in all cases $q_c < \tilde{q}_c$, indicating that these networks are less robust than predicted. When comparing the effects of each parameter, α has a much greater effect in reducing q_c than λ; sampling by distance plays a much greater role in determining robustness than how the network is constructed.

The curves in Fig. 2.6 show S_2 for the entire range of q; to study q_c requires examining the peaks of these curves. Figure 2.7 systematically summarizes q_c as a function of λ and α. Over all parameters, q_c ranges from approximately 0.30 to 0.50. Globally, the most vulnerable region is at **A** ($(\lambda, \alpha) \approx (0, 2)$); these non-spatial networks with strong, superlinear ($\alpha > 1$) failure dependence occupy the most vulnerable region of Fig. 2.7 since their construction (low distance dependence) is in direct opposition to how links fail (high distance dependence). Even when networks are built with the goal of minimizing physical distances along links (high λ), the

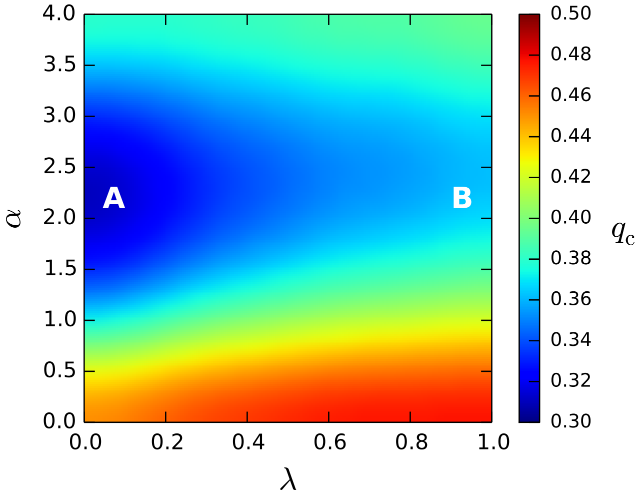
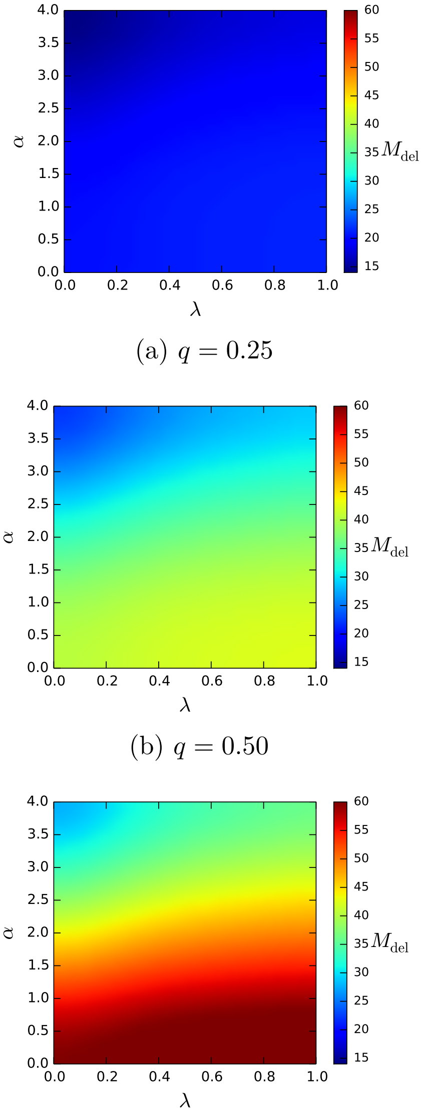


Figure 2.7: (Color online) Critical failure probability q_c as a function of α and λ. Overall, values of $\alpha > 0$ always correspond to lower robustness than when $\alpha = 0$ and in particular, the percolation threshold, q_c, is lowest near ***A*** *($(\lambda, \alpha) \approx (0, 2)$), while networks are generally most robust when $\alpha \leq 0.5$. The exponent α lowers q_c even in geometric networks (high λ) where spatial distance plays a stronger role in the network topology (region* ***B****). This matrix was smoothed with a $\sigma =$ 5-pixel gaussian convolution for clarity (1 pixel $= 0.025\ \lambda \times 0.1\ \alpha$).*

exponent α still lowers q_c compared with the theoretical prediction (highlighted at region **B**). Almost any introduction of spatial dependence on link failure (compare $\alpha > 0$ with $\alpha = 0$) leads to less robust networks.

Figure 2.7 also shows a slight increase in q_c for $\alpha > 2.5$. This is a result of Eq. (2.3): for such extreme values of α the failure probabilities Q_{ij} become concentrated and thus relatively fewer links are deleted for a given q, causing the apparent rise in q_c. This does not occur for $\alpha < 2.5$. We see this in Fig. 2.8 where we plot the number of deleted links (M_{del}) as a function of α and λ.

Finally, to better understand why these infrastructure networks are less robust than the theoretical prediction [37, 38], we studied correlations in network structure by computing the mean degree of nearest neighbors $\langle k_{\mathrm{nn}} \rangle = \sum_{k'} P(k' \mid k)$ [127] and the mean distance to nearest neighbors $\langle d_{\mathrm{nn}} \rangle = \int d' P(d' \mid k)\ dd'$, both as functions of node degree k. Here $P(k'|k)$ is the conditional probability that a node of degree k has a neighbor of degree k' and $P(d'|k)$ is the conditional probability that a node of degree k has a link of length between d' and $d' + dd'$. See Fig. 2.9. Due to the optimization (Eq. 2.1), both $\langle d_{\mathrm{nn}} \rangle$ and $\langle k_{\mathrm{nn}} \rangle$ indicate non-random structure, since they depend on k. Even for the case $\lambda = 1$, which shows no relationship between $\langle d_{\mathrm{nn}} \rangle$ and k, there is a positive trend for $\langle k_{\mathrm{nn}} \rangle$. Therefore, the optimized networks



(a) $q = 0.25$

(b) $q = 0.50$

(c) $q = 0.75$

Figure 2.8: (Color online) The number of deleted links M_{del} as a function of λ and α for several values of q. When $\alpha < 2$ and $q \leq 0.5$, M_{del} is almost exactly constant, but for larger α and q the number of deleted links begins to drop, as failure probabilities become "concentrated" on a smaller fraction of links. This causes the small rise on q_{c} for $\alpha > 2.5$ observed in Figure 2.7.

always possess correlated topologies.

Taken together, Fig. 2.9 shows that, beyond finite-size effects, $\tilde{q}_c$ overestimates q_c because (i) these networks are non-random and (ii) higher degree nodes tend to have longer links leading to hubs that suffer more damage when $\alpha > 0$. Since hubs play an outsized role in holding the network together, the positive correlation between d and k causes spatial networks to more easily fall apart, lowering their robustness.

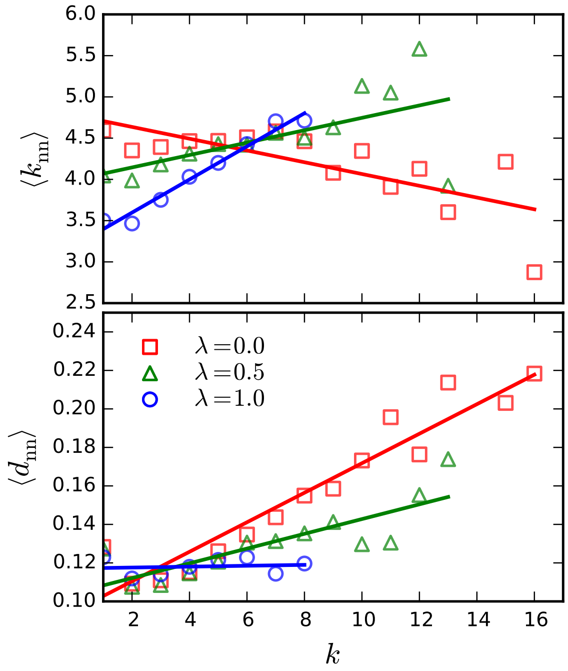


Figure 2.9: (Color online) Degree and distance correlations in optimized spatial networks. Here $\langle k_{nn} \rangle$ is the mean degree of nearest neighbors and $\langle d_{nn} \rangle$ is the mean distance of nearest neighbors. We observe that $\langle k_{nn} \rangle$ shows a negative trend with degree k for $\lambda = 0$, and positive trend for $\lambda = 0.5$ and 1.0. On the other hand, $\langle d_{nn} \rangle$ shows an increasing trend with k for decreasing λ. These optimized networks are not randomly constructed; they possess correlations in either network or spatial structure (or both) for all λ. The above metrics indicate that non-spatial networks form hubs whose longer links are likely to fail with higher probability and cause more damage to the network. Alternatively, more spatially-dependent networks (higher λ) have $\langle d_{nn} \rangle$ that depends less on k, indicating that link failures are spread somewhat more uniformly across high- and low-degree nodes.

2.5 Discussion

A potential application of this model is to designing microgrids. The microgrid concept, most commonly implemented in military settings, has gained wider popularity with the advent of the smart grid. Building a microgrid that is robust to failures while constrained by a budget is important for the widespread adoption of microgrids. Furthermore, the model also brings to light the need to keep in mind that the construction of convenient, long power lines may not be an optimal choice when accounting for the system's robustness. This may reinforce distributed generation across many buildings, as opposed to the power grid (traditional utility) creation of power lines stemming from a centralized cluster of small power plants. A move toward distributed generation and the decommissioning of the traditional utility may raise the overall stability of the grid. Existing infrastructure can use methods that reduce the power grid's dependence on distance (effectively lowering α), such as using towers to raise long-distance transmission lines above trees or otherwise protecting longer links. Distributed generation may be a cost-effective alternative.

Of course, the metrics used here are not all-encompassing for quantifying robustness. Additional measures may be used that go beyond the topological connectivity of networks to network functionality and dynamics, including problem-specific analyses [72, 39]. One specific example: it is worth understanding how a spatial network's travel distance may change following link failures, even when a giant component remains. It is also worth further characterizing fluctuations in, e.g., q_c that are due to the small size of these micronetworks.

Both the network construction budget and system size (number of nodes) were treated as constants in this work, for simplicity. Yet studying their interplay with the system's robustness may reveal important features of microgrids at different scales. Additional future work may include considering the unit square to have differential terrain, changing the cost of edge placement over a continuous gradient. Also, applying the existing model to real infrastructure network data, we may measure the robustness of critical networks and have better insight on how to design and improve these structures. Furthermore, in a real power grid nodes do not all have equal roles and thus investigating not only spatially-dependent edge failure but variations in node importance may gain more insight into spatial network robustness.

Acknowledgments

We thank M. T. Gastner, J. R. Williams, N. A. Allgaier, P. Rezaei, P. D. Hines, and P. S. Dodds for useful discussions. This research funded and supported by the National Science Foundation's IGERT program (DGE-1144388), Vermont Complex

Systems Center, and the Vermont Advanced Computing Core, which is supported by NASA (NNX 06AC88G).

2.6 Appendix

2.6.1 Constructing optimized networks

Networks are initialized by first placing $N = 50$ nodes uniformly at random inside the unit square. Initially the network is empty. The minimum spanning tree (MST) is inserted between these nodes using Kruskal's algorithm [97] with link weights corresponding to d_{eff}, and the construction cost and travel distance are computed. The spanning tree, which may be modified as optimization progresses, ensures the travel distance is finite when optimization begins.

We find solutions to the constrained optimization problem (Eq. (2.1)) using simulated annealing (SA). At the beginning of each SA step, an edge is added to the network at random and construction cost and travel distance are recomputed. If the budget constraint is still satisfied with the addition of this edge, the edge is kept using Boltzmann's criterion: the edge is retained if it lowers the travel distance; if it does not lower the travel distance it is retained with probability $e^{-\beta \Delta E}$, where ΔE is the change in travel distance due to this change in the network, and β acts as the inverse temperature.

If the random edge puts the network over budget, we remove it and do one of two modifications. With probability one half an existing edge is moved by placing it at random in the network where no edge exists. Otherwise, a rewire is chosen. Edges are rewired by first selecting an existing edge at random, next selecting either of the nodes connected by that edge, and finally attaching that end of the edge from the chosen node to a node that is a non-neighbor. In other words, edge (i, j) is removed and edge (i, k) is inserted where $k \neq j$ and k was not previously a neighbor of i. The move/rewire perturbation is then kept using the same Boltzmann's criterion.

The cooling schedule starts at $\beta_0 = 100/(\text{cost of MST})$, and cooled subsequently as $\beta_{t+1} = \beta_t \left(1 + 3 \times 10^{-5}\right)$. At each SA step we check if the current network topology is the best seen to that point; the most optimal network found during any of the 3×10^5 total SA steps is taken as our solution.

Chapter 3

Discovering perceptions via weighted networks

Abstract

Identifying and communicating relationships between causes and effects is important for understanding our world, but is affected by language structure, cognitive and emotional biases, and the properties of the communication medium. Despite the increasing importance of social media, much remains unknown about causal statements made online. To study real-world causal attribution, we extract a large-scale corpus of causal statements made on the Twitter social network platform as well as a comparable random control corpus. We compare causal and control statements using statistical language and sentiment analysis tools. We find that causal statements have a number of significant lexical and grammatical differences compared with controls and tend to be more negative in sentiment than controls. Causal statements made online tend to focus on news and current events, medicine and health, or interpersonal relationships, as shown by topic models. By quantifying the features and potential biases of causality communication, this study improves our understanding of the accuracy of information and opinions found online.

3.1 Introduction

Social media and online social networks now provide vast amounts of data on human online discourse and other activities [101, 10, 84, 126, 172, 43, 118, 152]. With so much communication taking place online and with social media being capable of hosting powerful misinformation campaigns [138] such as those claiming vaccines cause autism [146, 98], it is more important than ever to better understand the discourse of causality and the interplay between online communication and the statement of cause and effect.

Causal inference is a crucial way that humans comprehend the world, and it has been a major focus of philosophy, statistics, mathematics, psychology, and the cognitive sciences. Philosophers such as Hume and Kant have long argued whether causality is a human-centric illusion or the discovery of a priori truth [77, 83]. Causal inference in science is incredibly important, and researchers have developed statistical measures such as Granger causality [65], mathematical and probabilistic frameworks [143, 149, 56, 128], and text mining procedures [63, 129, 93] to better infer causal influence from data. In the cognitive sciences, the famous perception experiments of Michotte *et al.* led to a long line of research exploring the cognitive biases that humans possess when attempting to link cause and effect [140, 148, 81].

How humans understand and communicate cause and effect relationships is complicated, and is influenced by language structure [90, 161, 91, 70] and sentiment or valence [23]. A key finding is that the perceived emphasis or causal weight changes between the agent (the grammatical construct responsible for a cause) and the patient (the construct effected by the cause) depending on the types of verbs used to describe the cause and effect. Researchers have hypothesized [26] that this is because of the innate weighting property of the verbs in the English language that humans use to attribute causes and effects. Another finding is the role of a valence bias: the volume and intensity of causal reasoning may increase due to negative feedback or negative events [23].

Despite these long lines of research, causal attributions made via social media or online social networks have not been well studied. The goal of this paper is to explore the language and topics of causal statements in a large corpus of social media taken from Twitter. We hypothesize that language and sentiment biases play a significant role in these statements, and that tools from natural language processing and computational linguistics can be used to study them. We do not attempt to study the factual correctness of these statements or offer any degree of verification, nor do we exhaustively identify and extract all causal statements from these data. Instead, here we focus on statements that are with high certainty causal statements, with the goal to better understand key characteristics about causal statements that differ from everyday online communication.

The rest of this paper is organized as follows: In Sec. 4.2 we discuss our materials and methods, including the dataset we studied, how we preprocessed that data and extracted a ‘causal’ corpus and a corresponding ‘control’ corpus, and the details of the statistical and language analysis tools we studied these corpora with. In Sec. 4.4 we present results using these tools to compare the causal statements to control statements. We conclude with a discussion in Sec. 4.5.

3.2 Materials and Methods

Dataset, filtering, and corpus selection

Data was collected from a 10% uniform sample of Twitter posts made during 2013, specifically the Gardenhose API. Twitter activity consists of short posts called tweets which are limited to 140 characters. Retweets, where users repost a tweet to spread its content, were not considered. (The spread of causal statements will be considered in future work.) We considered only English-language tweets for this study. To avoid cross-language effects, we kept only tweets with a user-reported language of ‘English’ and, as a second constraint, individual tweets needed to match more English stopwords than any other language’s set of stopwords. Stopwords considered for each language were determined using NLTK’s database [21]. A tweet will be referred to as a ‘document’ for the rest of this work.

All document text was processed the same way. Punctuation, XML characters, and hyperlinks were removed, as were Twitter-specific “at-mentions” and “hashtags”. There is useful information here, but it is either not natural language text, or it is Twitter-specific, or both. Documents were broken into individual words (unigrams) on whitespace. Casing information was retained, as we will use it for our Named Entity analysis, but otherwise all words were considered lowercase only. Stemming [106] and lemmatization [131] were not performed.

Causal documents were chosen to contain one occurrence only of the exact unigrams: ‘caused’, ‘causing’, or ‘causes’. The word ‘cause’ was not included due to its use as a popular contraction for ‘because’. One ‘cause-word’ per document restricted the analysis to single relationships between two relata. Documents that contain *bidirectional* words (‘associate’, ‘relate’, ‘connect’, ‘correlate’, and any of their stems) were also not selected for analysis. This is because our focus is on causality, an inherently one-sided relationship between two objects. We also did not consider additional synonyms of these cause words, although that could be pursued for future work. *Control documents* were also selected. These documents did not contain any of ‘caused’, ‘causing’, or ‘causes’, nor any bidirectional words, and are further matched temporally to obtain the same number of control documents as causal documents in each fifteen-minute period during 2013. Control documents were otherwise selected randomly; causal synonyms may be present. The end result of this procedure identified 965,560 causal and 965,560 control documents. Each of the three "cause-words", 'caused', 'causes', and 'causing' appeared in 38.2%, 35.0%, and 26.8% of causal documents, respectively.

Tagging and corpus comparison

Documents were further studied by annotating their unigrams with **Parts-of-Speech** (POS) and **Named Entities** (NE) tags. POS tagging was done using NLTK v3.1 [21] which implements an averaged perceptron classifier [160] trained on the Brown Corpus [55]. POS tags denote the nouns, verbs, and other grammatical constructs present in a document. Named Entity Recognition (NER) was performed using the 4-class, distributional similarity tagger provided as part of the Stanford CoreNLP v3.6.0 toolkit [110]. NER aims to identify and classify proper words in a text. The NE classifications considered were: Organization, Location, Person, and Misc. The Stanford NER tagger uses a conditional random field model [54] trained on diverse sets of manually-tagged English-language data (CoNLL-2003) [110]. Conditional random fields allow dependencies between words so that 'New York' and 'New York Times', for example, are classified separately as a location and organization, respectively.

Comparing corpora Unigrams, POS, and NEs were compared between the cause and control corpora using **odds ratios** (ORs):

$$\mathrm{OR}(x) = \frac{p_C(x)/(1 - p_C(x))}{p_N(x)/(1 - p_N(x))}, \tag{3.1}$$

where $p_C(x)$ and $p_N(x)$ are the probabilities that a unigram, POS, or NE x occurs in the causal and control corpus, respectively. These probabilities were computed for each corpus separately as $p(x) = f(x)/\sum_{x' \in V} f(x')$, where $f(x)$ is the total number of occurrences of x in the corpus and V is the relevant set of unigrams, POS, or NEs. Confidence intervals for the ORs were computed using Wald's methodology [3].

As there are many unique unigrams in the text, when computing unigram ORs we focused on the most meaningful unigrams within each corpus by using the following filtering criteria: we considered only the ORs of the 1500 most frequent unigrams in that corpus that also have a term-frequency-inverse-document-frequency (tf-idf) score above the 90th percentile for that corpus [95]. The tf-idf was computed as

$$\text{tf-idf}(w) = \log f(w) \times \log\left(\frac{D}{df(w)}\right), \tag{3.2}$$

where D is the total number of documents in the corpus, and $df(w)$ is the number of documents in the corpus containing unigram w. Intuitively, unigrams with higher tf-idf scores appear frequently, but are not so frequent that they are ubiquitous through all documents. Filtering via tf-idf is standard practice in the information retrieval and data mining fields.

Cause-trees

For a better understanding of the higher-order language structure present in text phrases, *cause-trees* were constructed. A cause-tree starts with a root cause word (either 'caused', 'causing' or 'causes'), then the two most probable words following (preceding) the root are identified. Next, the root word plus one of the top probable words is combined into a bigram and the top two most probable words following (preceding) this bigram are found. Repeatedly applying this process builds a binary tree representing the n-grams that begin with (terminate at) the root word. This process can continue until a certain n-gram length is reached or until there are no more documents long enough to search.

Sentiment analysis

Sentimental analysis was applied to estimate the emotional content of documents. Two levels of analysis were used: a method where individual unigrams were given crowdsourced numeric sentiment scores, and a second method involving a trained classifier that can incorporate document-level phrase information.

For the first sentiment analysis, each unigram w was assigned a crowdsourced "labMT" sentiment score $s(w)$ [43]. (Unlike [43], scores were re-centered by subtracting the mean, $s(w) \leftarrow s(w) - \langle s \rangle$.) Unigrams determined by volunteer raters to have a negative emotional sentiment ('hate','death', etc.) have $s(w) < 0$, while unigrams determined to have a positive emotional sentiment ('love', 'happy', etc.) tend to have $s(w) > 0$. Unigrams that have labMT scores and are above the 90th percentile of tf-idf for the corpus form the set $\tilde{V}$. (Unigrams in $\tilde{V}$ need not be among the 1500 most frequent unigrams.) The set $\tilde{V}$ captures 87.9% (91.5%) of total unigrams in the causal (control) corpus. Crucially, the tf-idf filtering ensures that the words 'caused', 'causes', and 'causing', which have a slight negative sentiment, are not included and do not introduce a systematic bias when comparing the two corpora.

This sentiment measure works on a per-unigram basis, and is therefore best suited for large bodies of text, not short documents [43]. Instead of considering individual documents, the distributions of labMT scores over all unigrams for each corpus was used to compare the corpora. In addition, a **single sentiment score** for each corpus was computed as the average sentiment score over all unigrams in that corpus, weighed by unigram frequency: $\sum_{w\in\tilde{V}} f(w)s(w) \Big/ \sum_{w'\in\tilde{V}} f(w')$.

To supplement this sentiment analysis method, we applied a second method capable of estimating with reasonable accuracy the sentiment of individual documents. We used the sentiment classifier [156] included in the Stanford CoreNLP v3.6.0 toolkit to documents in each corpus. Documents were individually classified into one of five categories: very negative, negative, neutral, positive, very positive. The data used to train this classifier is taken from positive and negative reviews of movies (Stanford Sentiment Treebank v1.0) [156].

Topic modeling

Lastly, we applied topic modeling to the causal corpus to determine what are the topical foci most discussed in causal statements. Topics were built from the causal corpus using Latent Dirichlet Allocation (LDA) [22]. Under LDA each document is modeled as a bag-of-words or unordered collection of unigrams. Topics are considered as mixtures of unigrams by estimating conditional distributions over unigrams: $P(w|T)$, the probability of unigram w given topic T and documents are considered as mixtures of topics via $P(T|d)$, the probability of topic T given document d. These distributions are then found via statistical inference given the observed distributions of unigrams across documents. The total number of topics is a parameter chosen by the practitioner. For this study we used the MALLET v2.0.8RC3 topic modeling toolkit [113] for model inference. By inspecting the most probable unigrams per topic (according to $P(w|T)$), we found 10 topics provided meaningful and distinct topics.

3.3 Results

We have collected approximately 1M causal statements made on Twitter over the course of 2013, and for a control we gathered the same number of statements selected at random but controlling for time of year (see Methods). We applied **Parts-of-Speech** (POS) and **Named Entity** (NE) taggers to all these texts. Some post-processed and tagged example documents, both causal and control, are shown in Fig. 3.1A. We also applied sentiment analysis methods to these documents (Methods) and we have highlighted very positive and very negative words throughout Fig. 3.1.

In Fig. 3.1B we present odds ratios for how frequently unigrams (words), POS, or NE appear in causal documents relative to control documents. The three unigrams most strongly skewed towards causal documents were ‘stress’, ‘problems’, and ‘trouble’, while the three most skewed towards control documents were ‘photo’, ‘ready’, and ‘cute’. While these are only a small number of the unigrams present, this does imply a negative sentiment bias among causal statements (we return to this point shortly).

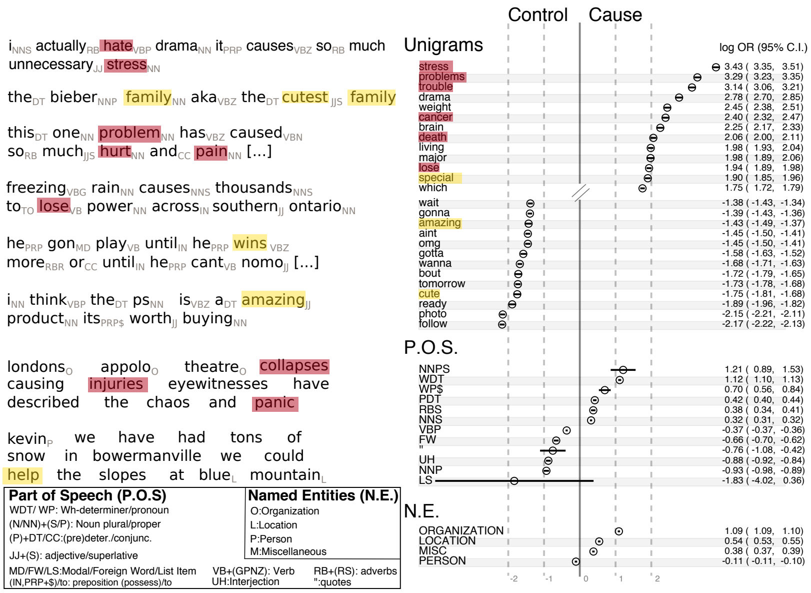


*Figure 3.1: Measuring the differences between causal and control documents. (**A**) Examples of processed documents tagged by Parts-of-Speech (POS) or Named Entities (NEs). Unigrams highlighted in red (yellow) are in the bottom 10% (top 10%) of the labMT sentiment scores. (**B**) Log Odds ratios with 95% Wald confidence intervals for the most heavily skewed unigrams, POS, and all NEs between the causal and control corpus. POS tags that are plural and use Wh-pronouns (that, what, which, ...) are more common in the causal corpus, while singular nouns and list items are more common in the controls. Finally, the ‘Person’ tag is the only NE less likely in the causal corpus. Certain unigrams were censored for presentation only, not analysis. All shown odds ratios were significant at the* $\alpha = 0.05$ *level except LS (List item markers).*

Figure 3.1B also presents odds ratios for POS tags, to help us measure the differences in grammatical structure between causal and control documents. The causal corpus showed greater odds for plural nouns (Penn Treebank tag: NNS), plural proper nouns (NNPS), Wh-determiners/pronouns (WDT, WP$) such as ‘whichever’,‘whatever’, ‘whose’, or ‘whosever’, and predeterminers (PDT) such as ‘all’ or ‘both’. Predeterminers quantify noun phrases such as ‘all’ in ‘after *all* the events that caused you tears’, showing that many causal statements, despite the potential brevity of social media, can encompass or delineate classes of agents and/or patients. On the other hand, the causal corpus has lower odds than the control corpus for list items (LS), proper singular nouns (NNP), and interjections (UH).

Lastly, Fig. 3.1B contains odds ratios for NE tags, allowing us to quantify the types of proper nouns that are more or less likely to appear in causal statements. Of the four tags, only the “Person” tag is less likely in the causal corpus than the control. (This matches the odds ratio for the proper singular noun discussed above.) Perhaps surprisingly, these results together imply that causal statements are less likely to involve individual persons than non-causal statements. There is considerable celebrity news and gossip on social media [172]; discussions of celebrities may not be especially focused on attributing causes to these celebrities. All other NE tags, Organization, Location, and Miscellaneous, occur more frequently in the causal corpus than the control. All the odds ratios in Fig. 3.1B were significant at the $\alpha = 0.05$ level except the List item marker (LS) POS tag.

The unigram analysis in Fig. 3.1 does not incorporate higher-order phrase structure present in written language. To explore these structures specifically in the causal corpus, we constructed “cause-trees”, shown in Fig. 3.2. Inspired by association mining [2], a cause-tree is a binary tree rooted at either ‘caused’, ‘causes’, or ‘causing’, that illustrates the most frequently occurring n-grams that either begin or end with that root cause word (see Methods for details).

The “causes” tree shows the focused writing (sentence segments) that many people use to express either the relationship between their own actions and a cause-and-effect (“even if it causes”), or the uncontrollable effect a cause may have on themselves: “causes me to have” shows a person’s inability to control a causal event (“[. . .] i have central heterochromia which causes me to have dual colors in both eyes”). The ‘causing’ tree reveals our ability to confine causal patterns to specific areas, and also our ability to be affected by others causal decisions. Phrases like “causing a scene in/at” and “causing a ruckus in/at” (from documents like “causing a ruckus in the hotel lobby typical [. . .]”) show people commonly associate bounds on where causal actions take place. The causing tree also shows people’s tendency to emphasize current negativity: Phrases like “pain this is causing” coming from documents like “cant you see the pain you are causing her” supports the sentiment bias that causal attribution is more likely for negative cause-effect associations. Finally, the ‘caused’ tree focuses heavily on negative events and indicates people are more likely to remember negative causal events. Documents with phrases from the caused tree (“[. . .] appalling tragedy [. . .] that caused the death”, “[. . .] live with this pain that you caused when i was so young [. . .]”) exemplify the negative events that are focused on are large-scale tragedies or very personal negative events in one’s life.

Taken together, the popularity of negative sentiment unigrams (Fig. 3.1) among causal documents shows that emotional sentiment or “valence” may play a role in how people perform causal attribution [23]. The “if it bleeds, it leads” mentality among news media, where violent and negative news are more heavily reported, may appeal to this innate causal association mechanism. (On the other hand, many news media

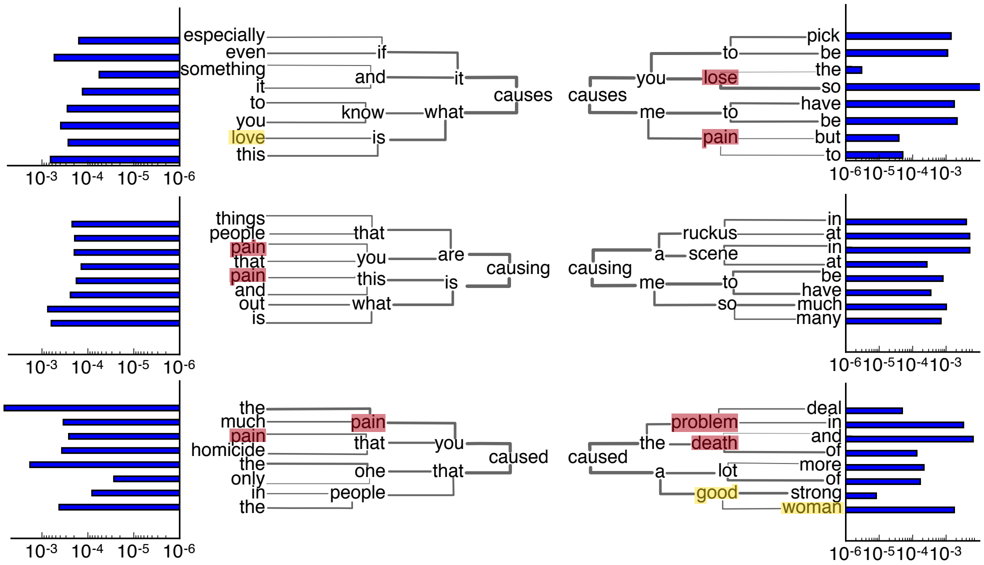


Figure 3.2: "Cause-trees" containing the most probable n-grams terminating at (left) or beginning with (right) a chosen root cause-word (see Methods). Line widths are log proportional to their corresponding n-gram frequency and bar plots measure the 4-gram per-document rate $N(4\text{-gram})/D$. Most trees express negative sentiment consistent with the unigram analysis (Fig. 3.1). The 'causes' tree shows (i) people think in terms of causal probability ("you know what causes [. . .]"), and (ii) people use causal language when they are directly affected or being affected by another ("causes you", "causes me"). The 'causing' tree is more global ("causing a ruckus/scene") and ego-centric ("pain you are causing"). The 'caused' tree focuses on negative sentiment and alludes to humans retaining negative causal thoughts in the past.

themselves use social media for reporting.) The prevalence of negative sentiment also contrasts with the "better angels of our nature" evidence of Pinker [130], illustrating one bias that shows why many find the results of Ref. [130] surprising.

Given this apparent sentiment skew, we further studied sentiment (Fig. 3.3). We compared the sentiment between the corpora in four different ways to investigate the observation (Fig. 3.1B that people focus more about negative concepts when they discuss causality. First, we computed the mean sentiment score of each corpus using crowdsourced "labMT" scores weighted by unigram frequency (see Methods). We also applied tf-idf filtering (Methods) to exclude very common words, including the three cause-words, from the mean sentiment score. The causal corpus text was slightly negative on average while the control corpus was slightly positive (Fig. 3.3A). The difference in mean sentiment score was significant (t-test: $p < 0.01$).

Second, we moved from the mean score to the distribution of sentiment across all (scored) unigrams in the causal and control corpora (Fig. 3.3B). The causal corpus

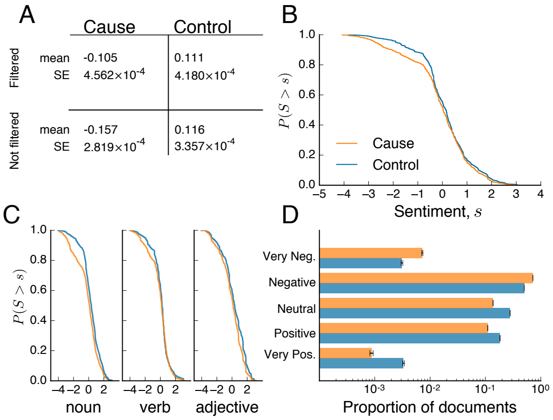


*Figure 3.3: Sentiment analysis revealed differences between the causal and control corpora. (**A**) The mean unigram sentiment score (see Methods), computed from crowdsourced "labMT" scores [43], was more negative for the causal corpus than for the control. This held whether or not tf-idf filtering was applied. (**B**) The distribution of unigram sentiment scores for the two corpora showed more negative unigrams (with scores in the approximate range $-3 < s < -1/2$) in the causal corpus compared with the control corpus. (**C**) Breaking the sentiment distribution down by Parts-of-Speech, nouns show the most pronounced difference in sentiment between cause and control; verbs and adjectives are also more negative in the causal corpus than the control but with less of a difference than nouns. POS tags corresponding to nouns, verbs, and adjectives together account for 87.8% and 77.2% of the causal and control corpus text, respectively. (**D**) Applying a different sentiment analysis tool—a trained sentiment classifier [156] that assigns individual documents to one of five categories—the causal corpus had an overabundance of negative sentiment documents and fewer positive sentiment documents than the control. This shift from very positive to very negative documents further supports the tendency for causal statements to be negative.*

contained a large group of negative sentiment unigrams, with labMT scores in the approximate range $-3 < s < -1/2$; the control corpus had significantly fewer unigrams in this score range.

Third, in Fig. 3.3C we used POS tags to categorize scored unigrams into nouns, verbs, and adjectives. Studying the distributions for each, we found that nouns explain much of the overall difference observed in Fig. 3.3B, with verbs showing a similar but smaller difference between the two corpora. Adjectives showed little difference. The distributions in Fig. 3.3C account for 87.8% of scored text in the causal corpus and 77.2% of the control corpus. The difference in sentiment between corpora was significant for all distributions (t-test: $p < 0.01$).

Fourth, to further confirm that the causal documents tend toward negative sentiment, we applied a separate, independent sentiment analysis using the Stanford NLP sentiment toolkit [156] to classify the sentiment of individual documents not unigrams (see Methods). Instead of a numeric sentiment score, this classifier assigns documents to one of five categories ranging from very negative to very positive. The classifier showed that the causal corpus contains more negative and very negative documents than the control corpus, while the control corpus contains more neutral, positive, and very positive documents (Fig. 3.3D).

We have found language (Fig. 3.1) and sentiment (Fig. 3.3) differences between causal statements made on social media compared with other social media statements. But *what* is being discussed? What are the topical foci of causal statements? To study this, for our last analysis we applied topic models to the causal statements. Topic modeling finds groups of related terms (unigrams) by considering similarities between how those terms co-occur across a set of documents.

We used the popular topic modeling method Latent Dirichlet Allocation (LDA) [22]. We ranked unigrams by how strongly associated they were with the topic. Inspecting these unigrams we found that a 10-topic model discovered meaningful topics. See Methods for full details. The top unigrams for each topic are shown in Tab. 3.1.

Topics in the causal corpus tend to fall into three main categories: (i) news, covering current events, weather, etc.; (ii) medicine and health, covering cancer, obesity, stress, etc.; and (iii) relationships, covering problems, stress, crisis, drama, sorry, etc. While the topics are quite different, they are all similar in their use of negative sentiment words.

3.4 Discussion

The power of online communication is the speed and ease with which information can be propagated by potentially any connected users. Yet these strengths come at a cost: rumors and misinformation also spread easily. Causal misattribution is at the heart of many rumors, conspiracy theories, and misinformation campaigns.

Given the central role of causal statements, further studies of the interplay of information propagation and online causal attributions are crucial. Are causal statements more likely to spread online and, if so, in which ways? What types of social media users are more or less likely to make causal statements? Will a user be more likely to make a causal statement if they have recently been exposed to one or more causal statements from other users?

The topics of causal statements also bring forth important questions to be addressed: how timely are causal statements? Are certain topics always being discussed in causal statements? Are there causal topics that are very popular for only brief periods and then forgotten? Temporal dynamics of causal statements are also interesting: do time-of-day or time-of-year factors play a role in how causal statements are made?

“News”	“Accident”	“Problems”	“Medical”	“Crisis”	“Sorry”	“Stress”	“Body”	“Drama”	“Injuries”
damage	traffic	dont	cancer	their	any	more	stress	like	his
fire	delays	people	break	our	never	than	lose	she	him
power	crash	they	some	from	been	being	weight	her	out
via	car	problems	men	how	sorry	over	stuff	lol	back
new	accident	why	can	about	there	person	living	out	her
news	death	about	disease	social	know	sleep	quickly	trouble	when
from	between	when	from	crisis	will	which	special	good	head
says	after	know	most	via	ive	people	proof	now	into
after	year	them	our	great	they	one	diets	sh*t	off
video	down	like	others	money	out	stress	excercise	life	well
global	there	drama	loss	many	problems	someone	f*ck	twitter	which
rain	man	who	heart	issues	can	makes	who	got	from
warming	due	one	health	should	now	think	giving	scene	down
water	from	youre	food	war	trouble	when	love	get	game
explosion	snow	get	symptoms	problems	see	most	god	too	fall
outage	road	stop	hair	government	one	thinking	people	girl	face
storm	old	think	women	true	how	brain	will	haha	then
change	over	how	blood	new	would	depression	our	needs	get
house	problems	sh*t	how	world	could	anxiety	those	see	injuries
may	chaos	want	skin	they	were	lack	one	walk	had
flooding	morning	cant	records	media	had	night	around	drama	sports
gas	two	because	adversity	other	ever	without	life	hes	stick
air	driving	too	high	obama	whats	love	his	woman	over
say	major	hate	helium	financial	again	mental	thats	some	while
stir	today	need	which	change	did	them	work	last	eyes
heavy	disruption	only	eating	violence	time	mind	out	strong	only
weather	train	really	may	will	think	fact	good	shes	hit
collapse	accidents	many	body	also	well	insomnia	sex	always	famous
climate	almost	even	smoking	shutdown	something	hand	come	him	hockey
death	into	then	own	issue	ill	even	great	ways	right
deaths	driver	someone	acne	support	still	feel	say	little	left
home	police	their	death	kids	about	physical	back	because	injury
oil	until	away	brain	problem	sure	emotional	when	said	got
massive	delay	always	alcohol	poor	hope	become	give	really	room
attack	congestion	feel	common	free	get	can	their	thats	involvement
blast	school	thats	deaths	says	youve	too	them	here	innumerable
two	late	say	news	pay	thats	less	things	man	time
city	weather	thing	treatment	against	day	same	goes	ass	play
into	been	something	unknown	party	some	often	comes	night	run
state	earlier	yourself	damage	confusion	good	keeps	too	ego	because

Table 3.1: Topical foci of causal documents. Each column lists the unigrams most highly associated (in descending order) with a topic, computed from a 10-topic Latent Dirichlet Allocation model. The topics generally fall into three broad categories: news, medicine, and relationships. Many topics place an emphasis on negative sentiment terms.

This analysis also shows the importance and potential predictive value sentiment plays in causal statements, with many causal statements skewed negative.

Our work here focused on a limited subset of causal statements, but more generally, these results may inform new methods for automatically detecting causal statements from unstructured, natural language text [63, 135]. Better computational tools focused on causal statements are an important step towards further understanding misinformation campaigns and other online activities. Lastly, an important but deeply challenging open question is how, if it is even possible, to validate the *accuracy* of causal statements. Can causal statements be ranked by some confidence metric(s)? We hope to pursue these and other questions in future research.

3.5 Appendix

3.5.1 Punctuation, Casing, and Parts-of-Speech

Parts-of-speech tagging depends on punctuation and casing, which we filtered in our data, so a study of how robust the POS algorithm is to punctuation and casing removal is important. We computed POS tags for the corpora with and without casing as well as with and without punctuation (which includes hashtags, links and at-symbols).

Two tags mentioned in Fig. 3.1B, NNPS and LS (which was not significant), were affected by punctuation removal. Otherwise, there is a strong correlation (Fig. 3.4) between Odds Ratios (causal vs. control) with punctuation and without punctuation, including casing and without casing ($\rho = 0.71$ and 0.80, respectively), indicating the POS differences between the corpora were primarily not due to the removal of punctuation or casing.

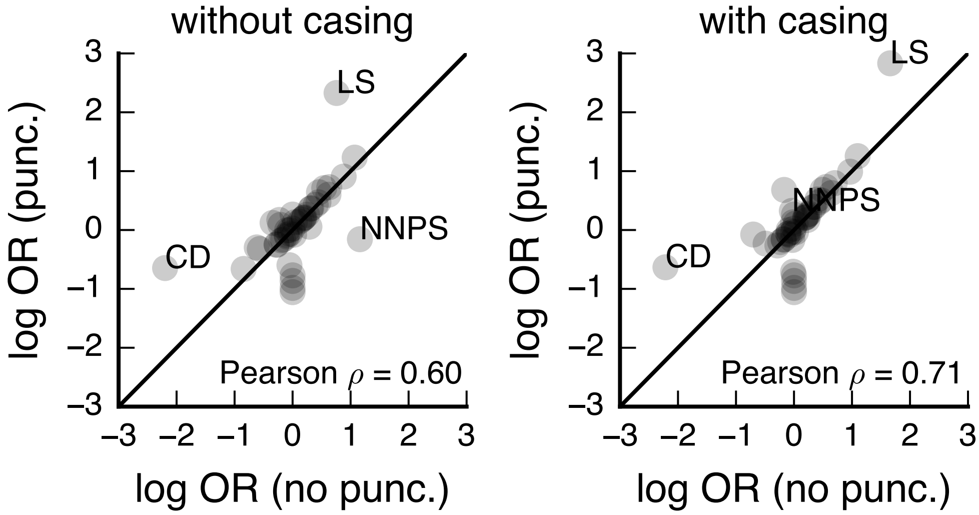


Figure 3.4: Comparison of Odds Ratios for all Parts-of-Speech (POS) tags with punctuation retained and removed for documents with and without casing. Tags Cardinal number (CD), List item marker (LS), and Proper noun plural (NNPS) were most affected by removing punctuation.

Chapter 4

Steering crowds via weighted networks

Abstract

Crowdsourcing is now invaluable in many domains for performing data collection and analysis by distributing tasks to workers, yet the true potential of crowdsourcing lies in workers not only performing tasks or answering questions but also in using their intuition and experience to contribute new tasks or questions for subsequent crowd analysis. Algorithms to efficiently assign tasks to workers focus on fixed question sets, but exploration of a growing set of questions presents greater challenges. For example, Markov Decision Processes made significant advances to question assignment algorithms, but they do not naturally account for hidden state transitions needed to represent newly contributed questions. We model growing question sets as growing networks of items linked by questions. If these networks grew at random they would obey classic 'rich-get-richer' dynamics, where the number of questions associated with an item depends on how early the item entered the network. This leads to more crowd time spent answering questions related to older items and less time exploring new items. We introduce a probability matching algorithm to curtail this bias by efficiently distributing workers between exploring new questions and addressing current questions. Experiments and simulations demonstrate that this algorithm can efficiently explore an unbounded set of questions while maintaining confidence in crowd answers.

4.1 Introduction and Related Work

The birth of the internet redefined almost every aspect of our lives, and recently this includes how we use human resources to get our work done. Crowdsourcing [75, 94, 25, 82] directs people who are available to complete tasks (workers) to others who need work completed (crowdsourcers). Crowdsourcing often distributes tasks that are easy for humans to solve, but may be difficult for a computer. Tasks are usually not given to a single worker, but completed by multiple workers and a statistical conclusion is drawn from their aggregated work. This takes advantage of multiple repeated answers to the same question, combining information to infer a probable answer [82]. For example, parsing human written text can be a difficult task and optical character recognition systems may be unable to identify all scanned words [108, 73, 86]. The reCaptcha [165] system takes scanned images of text which were difficult for computers to recognize and hands them off to internet workers for recognition. By solving quick and easy tasks, reCaptcha is able to translate massive quantities of text. Breaking a problem into manageable, parallel tasks lets workers finish tasks quickly while at the same time allows the crowdsourcer to maintain high confidence in aggregated work. Over the past years, different disciplines and companies have paid attention to the quality/efficiency benefits crowdsourcing offers, and an explosive boom of projects have been created [105, 137, 111].

Past research examines aggregation techniques in detail, but deciding on an optimal way to assign particular tasks to workers, and in what order, remains an active area of research. Task selection methods based on Thompson sampling [34] have been applied successfully to crowdsourcing [136, 1]. Previous work on optimal task assignment usually takes the form of a Markov Decision Process (MDP) [44, 79, 103]. MDP provides a rigorous mathematical framework to test policies for allocating jobs to workers [133, 134]. Several goals under this framework can be identified: optimal methods to aggregate answers [44, 78, 92, 88, 171] acknowledge that different workers possess unique expertise and use this to pair questions with workers who performed well in the past [141, 80]. A simple proposal to account for differing expertise weights worker's answers to a question relative to their historical performance on questions of a similar type [103].

Often a budget limits the total crowdsourcing resources available. [103, 85, 164, 163], either due to financial limits when workers are compensated or time constraints where the speed or size of the crowd are much smaller than the set of tasks to be performed or questions to be answered. Budgetary limit studies usually center on a few problem-specific scenarios such as the set of questions and worker's answers are fixed, or the set of questions are fixed and the worker's answers derive from a probability model. Likewise, algorithms may approach question selection sequentially and as a function of previous answers, or simultaneously. Sequential algorithms concentrate on picking questions to achieve higher reliability, but at the cost of a slower rate of answered questions. On the contrary, simultaneous algorithms sacrifice reliability for a high volume of answers.

To the best of our knowledge, previous studies considering efficient use of crowd resources do not consider a set of questions capable of growing as workers provide answers and also propose related questions. Yet, the truest expression of crowdsourcing must incorporate the intuition and experience of workers, who are potentially capable of providing the crowdsourcer with far more actionable information for many problem domains [24, 20]. To this end, we introduce a new type of question structure, a *question network*. As workers answer this single question they are given the opportunity to propose a related question and grow the network. The crowdsourcer must now leverage exploring the potential question space, and at the same time maintain a level of confidence for existing questions.

Study of question networks can be informed by the booming field of network science [51, 50, 14, 5, 169, 159, 120, 122]. Network science has studied concrete statistical properties governing how theoretical and real-world networks grow and behave. One property, the scale-free ('Rich-get-richer') degree distribution [14], is present in many real-world networks. In brief, a scale-free network contains a multitude of small-degree nodes and a handful of nodes with high-degree, connecting together large portions of the network.

This manuscript introduces a network perspective as a natural way to guide workers toward efficient exploration of a growing network of items and questions. In detail, *this manuscript makes the following novel contributions:*

1. The introduction of a growing network of linked questions with an accompanying theoretical analysis

2. The use of Thompson sampling to develop a crowd-steering algorithm that leverages efficient exploration of an evolving set of tasks or questions while maintaining confidence in answers.

3. Experiments that demonstrates the theoretical principles of a stochastic growing question network, and a second experiment that efficiently controls the network.

The rest of this paper is organized as follows: Section 4.2 describes and analyzes a network model of how a self-guided network is built from a crowd (Sec. 4.2.1), and Sec. 4.2.2 proposes a method for efficiently assigning questions to workers as the question set grows. Section 4.3 describes experiments to test the proposed theory and methods, Sec. 4.4 presents the results of these experiments, and Sec. 4.5 concludes with a discussion and future work.

4.2 Methods

We introduce a graphical model of a growing question network and study its properties under a null condition where the crowdsourcer assigns questions to workers randomly without use of a "steering" algorithm to provide guidance (Sec. 4.2.1). We then use these properties to develop a probability matching algorithm which provides said guidance to the crowdsourcer (Sec. 4.2.2).

4.2.1 Growing question network model

The upcoming methodology relies on the common language of network science, and the authors further modify this common language to more closely align with the application of crowdsourcing. We model a growing set of questions as a graph where vertices are items (the possible responses to a question) and edges are questions. A question network G is composed of a set of nodes and links (V, E), where $|V| = N$ and $|E| = M$. Edge weights record the answers given by workers. Those workers may also propose new questions (i.e., new combinations of new or existing items), leading to new vertices and edges. For example, consider a *synonymy proposal task* where workers are asked if two words u and v are synonyms. The question is the link (u, v) between two items representing those words. The worker may also be asked to propose another word w which is a synonym for u, v, or both words. This grows the question network by introducing new questions linking items (u, w), (v, w), or both. The degree k_i of item i counts the number of questions linked from this item to others.

We focus on the case where questions are binary, e.g, when workers are asked whether or not a link between two items should exist. Edge weights on links capture the number of 'yes' and 'no' answers given by workers. However, this graph representation is flexible enough to allow edge weights to contain any number of dimensions and there are no restrictions imposed on how workers propose questions. Moreover, this framework does not require such a graphical structure between items. It is capable of representing growing question sets without such relations, for example, a collection of disjoint questions always containing the response items 'True' and 'False' only.

Null model We propose a generative null model for a growing question network [15, 12]. Beginning from a network with one question, a crowdsourcer randomly chooses existing questions to send to workers also chosen at random. Those workers answer the questions and then with some probability also propose new questions. We study the properties of the network under these assumptions to motivate the development of a probability matching algorithm that can allow a crowdsourcer to efficiently explore the growing question network.

The null model initializes (at time $t = 0$) a network of 2 nodes with a single undirected and weighted link connecting them. A weight on each link tallies the number of times workers answered that the link should or should not exist. Under the null model, every link (i, j) has an associated *innovation rate* ρ_{ij}. The innovation rate defines the probability a random worker will introduce a new question into the network when presented with that particular question. If she chooses to innovate, the new question may relate to either or both of the items of the original question the worker was given.

Specifically, suppose a a random worker is given question (u, v) relating items u and v. Under the null model:

1. The worker answers question (u, v) (probability 1[1]).
2. The worker proposes a new item w to study (probability ρ_{uv}):
   (a) w is linked to one of the items of the original question (probability γ_{uv}). A single new question, either (u, w) or (v, w) chosen uniformly at random, is introduced.
   (b) w links to both items of the original question (probability $1 - \gamma_{uv}$). Two new questions, (u, w) and (v, w), are introduced.
3. Repeat from (1) with another sampled question and worker until termination.

See Fig. 4.1

Assuming different parameter values for each link is a simplification in that it assumes workers have comparable innovation rates. However, given sufficient data, a crowdsourcer can propose a statistical model for link parameters as well as for features of the workers, and use statistical inference to estimate these parameters during crowdsourcing (see also Discussion).

We now prove several *average* properties of this null model. The network's global properties explain the overall growth of the network, while local properties reveal how and when specific items gather questions. Studying the characteristics of the randomly growing, uncontrolled network informs policies that a crowdsourcer may use to manipulate the network (such as the algorithm we develop in Sec. 4.2.2).

The first theorem describes question growth in the random uncontrolled network.

Theorem 1 (Rate of question growth)**.** *The total number of links $M(t)$ as a function of time t can be modeled on average as $M(t) = \eta t + 1$ where $\eta = \langle\rho\rangle (2 - \langle\gamma\rangle)$ is termed the* ***exploration rate****.*

[1] We can easily incorporate worker dropout by supposing the crowdsourcer keeps selecting random workers until the first answer is given, but for simplicity we assume it is negligible here.

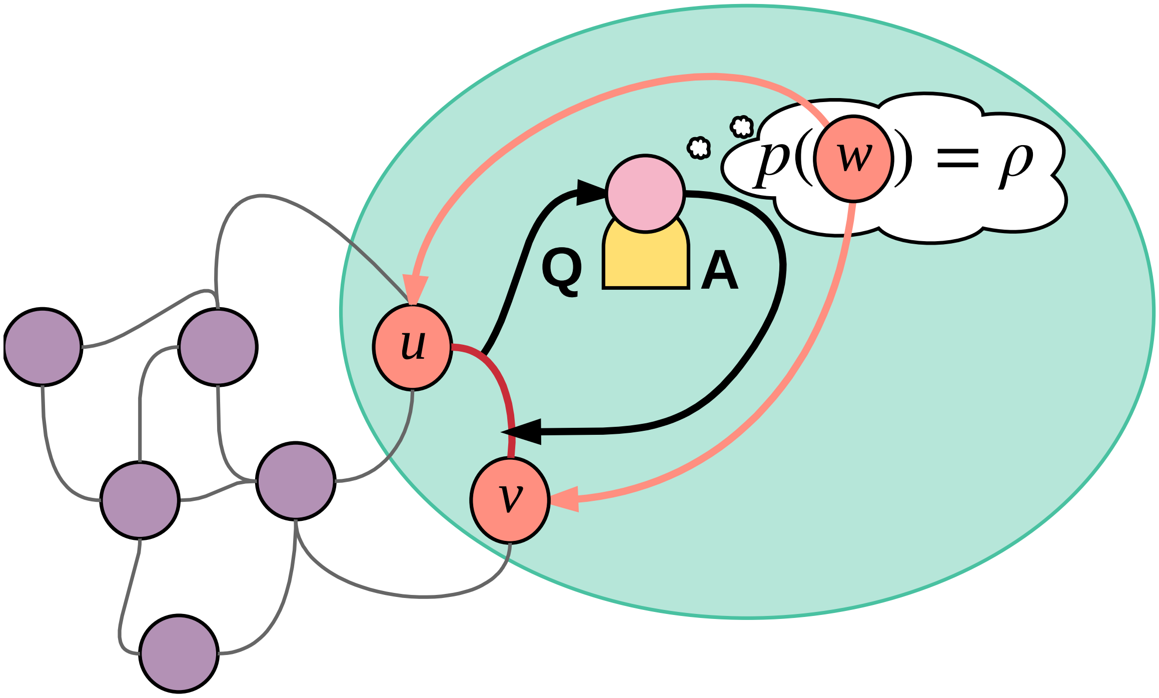


Figure 4.1: Interaction between the crowdsourcer posing a question (u, v) to a worker, the worker answering the question, and the worker innovating with a new item w with probability ρ (via the null model).

Proof. In order for the network to grow a worker must suggest an additional question, which occurs with probability on average $\langle \rho \rangle$ (average of ρ_{ij}). Once the worker commits to a suggestion, one question is added with probability on average $\langle \gamma \rangle$ or two questions are added with probability on average $1 - \langle \gamma \rangle$. Combining these two possibilities, the total number of questions grows on average over one timestep according to

$$M(t+1) = M(t) + \langle \gamma \rangle \langle \rho \rangle + 2 \langle \rho \rangle (1 - \langle \gamma \rangle),$$

with initial condition $M(0) = 1$ representing the single seed question of the network. Simplifying and making a continuum approximation, this difference equation becomes:

$$\frac{dM(t)}{dt} = \langle \rho \rangle (2 - \langle \gamma \rangle).$$

The exploration rate constant $\langle \rho \rangle (2 - \langle \gamma \rangle) \equiv \eta$ plays an important role in the overall network growth. This first-order ordinary differential equation has solution

$$M(t) = \eta t + 1. \tag{4.1}$$

□

The number of links grows linearly with a rate η that combines the average rates $\langle \rho \rangle$ and $\langle \gamma \rangle$. Intuitively, the network grows faster if questions are more likely to be innovative (larger $\langle \rho \rangle$), and/or the worker is able to suggest a question for both items at the same time (smaller $\langle \gamma \rangle$).

The solution to the rate equation for question growth can be used to compute the mean number of worker answers per question:

Theorem 2 (Mean answer density)**.** *The mean answer density (number of answers per question) is $\langle \mathcal{A} \rangle \to 1/\eta$ as $t \to \infty$.*

Proof. We define the mean number of answers per question as

$$\langle \mathcal{A} \rangle = \frac{\text{total number of answers}}{\text{total number of questions}}. \tag{4.2}$$

At every time step a question in the network accumulates a single answer from a worker. The denominator of (4.2) is the solution (4.1), and so the average density of answers per question must grow like

$$\langle \mathcal{A} \rangle = \frac{t}{\eta t + 1} = \frac{1}{\eta + \frac{1}{t}} \to \frac{1}{\eta}$$

as $t \to \infty$. □

The mean answer density represents the uncertainty in the system since there is generally more certainty (but not necessarily correctness) in crowd responses when more workers on average have independently answered questions. Controlling the answer density, and therefore the certainty, now boils down to controlling the exploration rate η. The mean answer density's dependence on η also encapsulates an 'exploration-exploitation' tradeoff: lower η leads to higher answer density, but at the cost of less exploration in the network; higher η increases the exploration but lowers answer density and makes more uncertainty in the network. In this null model, the crowdsourcer does not make choices that can exploit this, but tuning between these poles is a key component of the probability matching algorithm we introduce in Sec. 4.2.2.

The previous two theorems govern global properties of random question networks. We now turn to properties of individual items within the network to explain the unequal distribution of questions attached to items:

Theorem 3 (Rich-get-Richer model)**.** *A node i entering the network at time t_i will gain degree, on average, as $k_i(t) = \frac{\eta}{\langle\rho\rangle}\left(\frac{1+\eta t}{1+\eta t_i}\right)^{1/2}\mathcal{H}(t-t_i)$, where $\mathcal{H}$ is the Heaviside function.*

Proof. An existing item i only gains a question when the crowdsourcer chooses a question attached to i and the worker answering that question proposes a new question involving i. A question (i, j) associated with item i is selected by the crowdsourcer with probability $k_i(t)/M(t)$, where $k_i(t)$ is the degree (number of questions) of i at time t. After the worker answers question (i, j) she must innovate (probability $\langle\rho\rangle$) with an item w that is not already a neighbor of i (and $w \neq i$) and the new question must be (w, i) (probability $\langle\gamma\rangle/2$) or it must be two questions (w, i) and (w, j) (probability $1-\langle\gamma\rangle$). If the worker introduces question (w, j) only (probability $\langle\gamma\rangle/2$) then i does not gain a new question and so this possibility does not contribute to $k_i(t)$. Combining these possibilities together, $k_i(t)$ evolves on average according to

$$k_i(t) = k_i(t-1) + k_i\frac{\langle\rho\rangle}{M(t-1)}\left(\frac{\langle\gamma\rangle}{2} + (1-\langle\gamma\rangle)\right). \tag{4.3}$$

We approximate and simplify this difference equation as before:

$$\begin{aligned}\frac{dk_i}{dt} &= k_i\frac{\langle\rho\rangle}{M(t)}\left(1-\frac{\langle\gamma\rangle}{2}\right)\\ &= \frac{k_i}{2}\left(\frac{\eta}{\eta t+1}\right);\ k_i(t_i) = \frac{\eta}{\langle\rho\rangle},\end{aligned} \tag{4.4}$$

where $k_i(t_i)$ is the initial degree when item i was introduced at some time t_i. Integrating Eq. (4.4)

$$\int\frac{dk_i}{k_i} = \frac{1}{2}\int\frac{\eta}{\eta t+1}\,dt \tag{4.5}$$

results in

$$k_i(t) = \frac{\eta}{\langle\rho\rangle}\left(\frac{1+\eta t}{1+\eta t_i}\right)^{1/2}\mathcal{H}(t-t_i). \tag{4.6}$$

□

We see from this derivation that the rich-get-richer, preferential attachment mechanism [14] is automatic when questions are chosen at random: an item i is more likely to appear in a chosen question the more questions it has, and therefore items with more questions are more likely to gain further questions than other items. Further, the degree of an item depends critically on two parameters. The ratio of exploration rate η to $\langle\rho\rangle$ equally affects all items in the network. On the other hand, the time of entry t_i dampens the growth of items that enter the network late and increases the growth of earlier items. This phenomena is often referred to as the 'first mover's advantage', and in the context of crowdsourcing a growing network, items entered earlier in the system accrue more questions than later items.

Using the local estimate of item degree to derive the global degree distribution of the network, we find

Theorem 4 (Degree Distribution)**.** *The degree distribution of the growing question network*

$$P(k(t)) \to 2\left(\frac{\eta}{\langle\rho\rangle}\right)^2\frac{1}{k^3} \tag{4.7}$$

as $t \to \infty$.

Proof. Following [15]:

$$P(k_i(t) < k) = P\left(\frac{\eta}{\langle\rho\rangle}\left(\frac{1+\eta t}{1+\eta t_i}\right)^{1/2} < k\right) \tag{4.8}$$

defines an item i's cumulative distribution function. Solving (4.8) for the time of entry t_i we arrive at

$$P\left(\eta\left(\frac{1}{k\,\langle\rho\rangle}\right)^2(1+\eta t) - \frac{1}{\eta} < t_i\right)$$

and rearranging:

$$1 - P\left(t_i < \eta\left(\frac{1}{k\,\langle\rho\rangle}\right)^2(1+\eta t) - \frac{1}{\eta}\right). \tag{4.9}$$

Entry times t_i of items into the network follow a distribution proportional to $\langle\rho\rangle$ uniformly through time.

$$P(t_i = t) \propto \langle\rho\rangle\,,$$

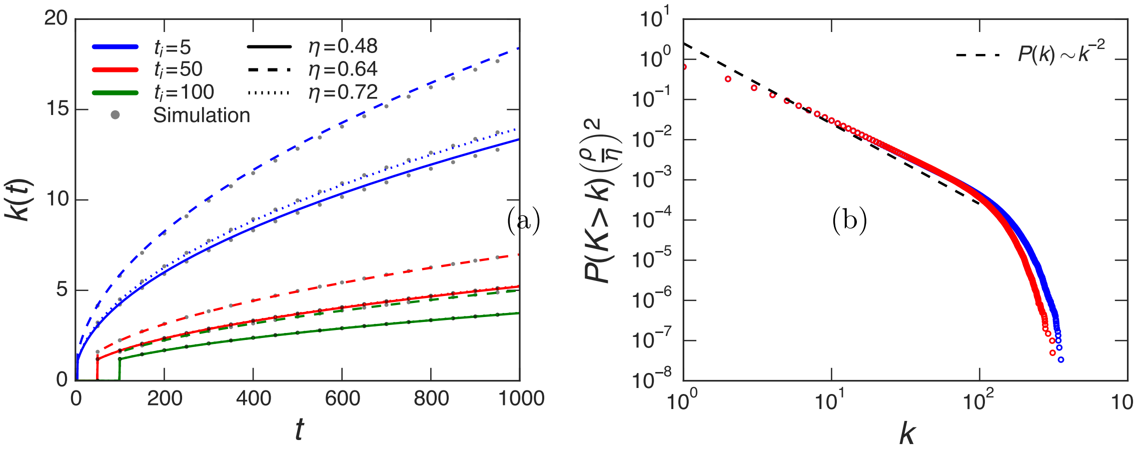


Figure 4.2: Agreement of theoretical predictions of network growth with simulations for several different choices of parameters.

and after normalizing we discover the time of entry follows a uniform distribution. Referring back to (4.9) and taking the integral definition of a cumulative distribution,

$$\begin{aligned} 1 - P\left(t_i < \eta\left(\frac{1}{k\left\langle\rho\right\rangle}\right)^2 (1+\eta t) - \frac{1}{\eta}\right) = \\ 1 - \frac{1}{t}\left(\frac{1}{k\left\langle\rho\right\rangle}\right)^2 \eta\left(1+\eta t\right) - \frac{1}{\eta t}\, dt. \end{aligned} \tag{4.10}$$

Differentiating (4.10) with respect to k uncovers the degree distribution:

$$\frac{\partial P(k_i < k)}{\partial k} = P(k(t)) = \frac{2\eta\left(\eta t + 1\right)}{t\left\langle\rho\right\rangle^2}\frac{1}{k^3} \to 2\left(\frac{\eta}{\left\langle\rho\right\rangle}\right)^2 \frac{1}{k^3} \tag{4.11}$$

as $t \to \infty$. □

Our theoretical analysis is supported by simulations of growing question networks. We conducted $5,000$ simulations and recorded the degree distribution $P(k)$ and degree k of items across different values of exploration rate η and time of item entry t_i. Figure 4.2(a) validates the slower rate of question accrual for late arriving items, and Fig. 4.2(b) shows the degree distribution's match to theory by the collapse of each curve over multiple values of η.

4.2.2 Probability matching algorithm for growing question sets and nets

Most algorithms for steering workers toward tasks choose questions by defining a metric that captures important characteristics in the system. For example, algorithms stressing accuracy often build metrics that reward higher numbers of answers for questions, achieving a p-value below a pre-defined threshold, or diminishing the variance of questions.

The framework of probability matching, specifically Thompson sampling [34] (TS), is one of the most powerful ways to efficiently choose from a set of dynamic options when choices must be made with limited information. Unlike greedy algorithms, one of the strengths of TS is that its stochastic nature prevents choosing locally optimal questions only.

To Thompson sample from a set of options, one assumes a random variable X which follows a distribution $\varphi(x \mid \theta_i(t))$, where $\theta_i(t)$ is a set of parameters specific to i at time t. One draws an $x_i(t)$ for each option i and selects the option j with the smallest x (or largest x, depending on what x represents), $j = \arg\min_i x_i(t)$. After option j is played (the worker's answer is received), the parameters for option j are updated. Often x is a Bernoulli random variable and it is natural for φ to be the conjugate Beta distribution with parameters α, β which are updated depending on whether $x = 0$ or $x = 1$.

For specific problems, TS depends on an appropriate reward function. In the context of crowdsourcing, one generally cannot verify the accuracy of crowd answers, so the best choice is to reward certainty or consensus. If the crowd is consistent in their responses for a given question, then that implies the question is being answered as well as possible under current conditions. Thus, in contrast to the Bernoulli Bandit problems typically studied with TS, we do not want to reward 'yes' answers over 'no' answers only. Instead, we want to reward choices that lower the crowdsourcer's measure of uncertainty for questions.

A natural measure of uncertainty for a categorical random variable is the Shannon entropy. However, efficiency is important to a crowdsourcer. If a question has 200 responses which are evenly split, that is very different than a question with 2 responses which is also evenly split, despite having the same entropy. Generally, the crowdsourcer would prefer to assign a worker to the latter question, as there is greater hope of lowering its uncertainty and the crowdsourcer is not spending further resources on a question which is unlikely to be informative.

This argument guides us to choosing a metric involving both the total number of answers to a question and how evenly distributed those answers were over the categories of that question. We introduce a metric called *link bias* (d) that is sensitive to the uncertainty of a question, but unlike entropy, also accounts for the total number of answers. To begin, the multinomial distribution, with $C-1$ parameters, naturally models the distribution of a general question's total number of answers T across C possible outcomes, and the Dirichlet distribution, conjugate to the multinomial, can estimate the parameters of the multinomial. Since we expect no available prior information, a non-informative prior can be used. In the case of two categories, which we focus on, the Dirichlet distribution reduces to the Beta distribution ($B(\alpha, \beta)$).

In order to define question uncertainty, we need a reference point. At a question's peak uncertainty, workers have answered evenly among the question's (C) categories causing an equal proportion of answers per category. In our binary case ($C = 2$), this corresponds to a proportion of $1/C = 1/2$. We transform the proportion of answers for question (i, j) to the distance from maximum uncertainty with $d \equiv \left|\frac{1}{2} - p_{ij}(1)\right|$, where $p_{ij}(1)$ is the fraction of '1' or 'yes' or 'true' answers. When $p_{ij} \sim B(\alpha, \beta)$, the probability density of d becomes

$$\varphi(d \,|\, \alpha, \beta) = \frac{(1-2d)^{\alpha-1}(1+2d)^{\beta-1} + (1+2d)^{\alpha-1}(1-2d)^{\beta-1}}{B\left(\alpha, \beta\right) 2^{\alpha+\beta-2}}, \tag{4.12}$$

where for simplicity the dependence of α, β on (i, j) has been suppressed. Intuitively, a low link bias ($d \approx 0$) is given to questions undecided by the majority yet (a uniform proportion of answers in each category), while a high link bias (at most $d = 1/2$) tells us the link was decided unanimously.

However, the link bias alone may not sufficiently steer the crowdsourcer to choose questions with a lower number of answers. If needed, we can combine a preference for sampling questions with a low number of answers, with a preference for question that are uncertain, by weighting (4.12) by the current number of answers to define a new 'weighted phi' metric φ_N:

$$\varphi_N(d \mid \alpha, \beta) \equiv \frac{N_{ij}\varphi(d \,|\alpha, \beta)}{\sum_{uv \in E} N_{uv}} \tag{4.13}$$

where N_{ij} is the total number of answers to question (i, j) at the time of sampling.

Thompson sampling of questions via φ or via φ_N defines the two probability matching algorithms we propose. These algorithms handle growing networks of questions automatically and are general enough to be applicable for problems without graphical relations between questions. We will conduct experiments on growing question networks testing the relative performance of both algorithms, and comparing them to other baseline strategies, such as randomly choosing questions.

4.3 Experiments

We conducted two experiments to test the theoretical analysis and the Thompson sampling methods. For the first experiment, we superimposed two distinct network structures onto a previously conducted crowdsourcing task [103] where questions have been time-ordered to mimic a growing question network, and used this to test three different question selection algorithms. The second experiment used a new *synonymy proposal task* we performed on the Mechanical Turk crowdsourcing platform [94], and exemplifies a true growing question network.

4.3.1 Experiment 1

In order to determine the effectiveness of choosing questions based on link bias, we first performed a four-armed experiment using the Recognizing Textual Entailment (RTE) dataset [103], a set of $8,000$ binary answers (0 or 1) over a set of 800 unique questions.

For simulating question growth, we superimposed graph structures onto the question set to link the 800 questions together. We built $5,000$ Erdős-Rényi (ER) and Barabasi-Albert (BA) networks [121]. These two options represent the extremes of network structure, and were chosen to test the robustness of question selection algorithms over different networks. An ER network [51] (specifically the $G(n,m)$ formulation) starts with a set of N nodes and 0 links; a pre-specified number of links M are placed in the network choosing randomly without replacement from all possible $\binom{N}{2}$ pairs of nodes. In contrast, the BA network [14] starts with 2 nodes joined by a single link, nodes are added one at a time until all N nodes are placed, and each new node attaches to m_0 existing nodes in the network. New nodes attach to an existing node i with probability $k_i / \sum_{n \in N} k_n$, a mechanism that is often called *preferential attachment*.

For simulation purposes, each ER network realization must contain exactly 400 nodes, 800 links, and be connected. BA networks are connected by design; we still enforced the same number of nodes (400) and links (800) as the ER networks. Each simulated crowdsourcing was initialized with one question (a link in the network connecting two corresponding item) chosen at random from the underlying network. During the simulated crowdsourcing, workers answer a question with a 1 with probability equal to the proportion of 1's observed in the original RTE dataset for that question, otherwise the worker answers 0. Next, and with probability $\langle \rho \rangle$, a new node (item) is introduced into the network by selecting randomly from the unseen neighbors of either i or j within that simulation's graph[2]. If there are no new items to

[2] This differs slightly from the analytic null model because there is no $\langle \gamma \rangle$. Instead, two links are

add corresponding to the selected question, this iteration is undone and the algorithm continues. All simulations were run with $\langle\rho\rangle = 0.20$ and $6,000$ time steps.

Simulations were performed independently for each of four arms. The condition of each arm governs how questions are selected by the simulated crowdsourcer:

Random: The first arm of the experiment had a condition where questions were chosen randomly from the pool of already visited edges.

Looping: The second arm uses a *looping* question selection algorithm. The first edge that entered the system is answered by a worker, then the second edge in the system is given to a worker, then the third edge and so on. When the algorithm reaches the most recent edge entered into the system it starts again from the oldest edge.

Thompson sampling with φ: The third arm uses Thompson sampling to select edges based on smallest link bias (φ).

Thompson sampling with φ_N: As in the third arm but links are Thompson sampled with φ_N instead of φ.

This experiment can demonstrate the strengths and weaknesses of selecting links based on φ_N versus random, looping, and φ edge selection. Results of Experiment 1 are presented in Sec. 4.4 and Fig. 4.4.

4.3.2 Experiment 2

This double-armed experiment created a growing question network from scratch, and evaluated the φ_N-based sampling versus random sampling. We paid workers on Amazon's Mechanical Turk crowdsourcing platform [94, 27] to participate in a **synonymy validation and proposal experiment**. Synonymy proposal is a good test application for the frameworks we study because workers can easily understand the task and are likely to be capable of proposing new questions (by suggesting new synonyms). Of course, data on synonymy relations are available in lexical resources such as WordNet [116], which may be used in this specific task for assessing the accuracy of proposed synonyms, but our goal with this task is validation and comparison of the frameworks. In Experiment 2, each worker completes up to 100 synonymy tasks being compensated at \$0.04 USD per task. Each synonymy task gives a pair of words to a worker and asks whether or not they are synonyms. After they answer either 'yes' or 'no', we allow the worker to suggest additional synonyms for each word of the given pair, or a single synonym associated with the combined word pair (Fig. 4.3).

formed automatically if the newly introduced item is linked to both i and j in the imposed network.

The question selection algorithm draws from all previous worker suggestions, and delivers a question to the next queued worker. The random arm chooses links using the same methodology as the random arm from Experiment 1, while the second arm selects links according to Thompson sampling of φ_N. Results for Experiment 2 are presented in Sec. 4.4 and Fig. 4.5.

4.3.3 Evaluation metrics

For the first experiment, we measure five attributes to decide the superior question selection algorithm. At each time step t, for each simulated network we record network properties

$$f_{\text{nodes}} = \sum_{u \in V(t)} \mathbb{1} / \sum_{v \in V(\infty)} \mathbb{1} \tag{4.14}$$

and

$$f_{\text{edges}} = \sum_{p \in E(t)} \mathbb{1} / \sum_{q \in E(\infty)} \mathbb{1}, \tag{4.15}$$

the fraction of items and the fraction of questions seen at time t, respectively, where $V(t)$ ($E(t)$) denote the number of items (questions) at time t, $V(\infty)$ ($E(\infty)$) denote the total number of items (questions) at the end of the experiment, and $\mathbb{1}$ is the indicator function.

Next, we record the entropy S and expected link bias $\bar{d}$ averaged over all currently visible questions to quantify uncertainty in the network:

$$S = -\sum_{ij \in E(t)} \sum_{x \in \{0,1\}} p_{ij}(x) \log_2 p_{ij}(x) \Big/ \sum_{ij \in E(t)} \mathbb{1}, \tag{4.16}$$

and

$$\bar{d} = \sum_{ij \in E(t)} \left| \frac{1}{2} - p_{ij}(1) \right| \Big/ \sum_{ij \in E(t)} \mathbb{1}, \tag{4.17}$$

where $p_{ij}(x)$ is the (laplace-smoothed) fraction of binary answers of x for question (i, j).

The final evaluation metric, mean answer density, measures how well covered each question is in a particular network (see also Thm. 2):

$$\mathcal{A} = \left(\sum_{ij \in E(t)} \sum_{x \in \{0,1\}} N_{ij}(x) \right) \Big/ \sum_{ij \in E(t)} \mathbb{1}, \tag{4.18}$$

where the $N_{ij}(x)$ represents the count of answer x corresponding to link (i, j) (at time t).

4.4 Results

Figure 4.4 displays the five evaluation metrics associated with Experiment 1, averaged over the $5,000$ ER and BA networks. We denote this average with $\langle \cdot \rangle$. The φ_N selection algorithm outperforms all others in exploration and uncertainty metrics across ER and BA networks. This selection algorithm explored more of the network, and faster, as evidenced by $\langle f_{\text{edges}} \rangle$ and $\langle f_{\text{nodes}} \rangle$, although all methods perform well along these measures. A lower $\langle S \rangle$ and higher $\langle \varphi \rangle$ compared with the other algorithms showed φ_N-based sampling suppresses uncertainty. Although the φ_N algorithm performed well on these metrics, $\langle \mathcal{A} \rangle$ fell below other algorithms in the BA network. The overall performance of φ_N-sampling in experimental simulation nominates it as an ideal candidate for Experiment 2's more realistic setting.

Experiment 2's random and probability sampled arms ended with $8,000$ and $7,840$ answers provided by 279 and 324 workers, respectively. Figure 4.5 shows the constructed random and sampled synonym networks. Qualitatively, we see that items are more evenly connected by questions in the probability sampled arm than in the random arm (Fig. 4.5(a) vs. Fig. 4.5(b)).

Quantitatively, in Fig. 4.5(c) the probability sampling algorithm discovered more questions and items than the random algorithm (1963 and 927 versus 3103 and 1464), while only sacrificing a marginal amount of certainty (measured by $\langle S \rangle$ or $\left\langle \vec{d} \right\rangle$) in the questions proposed (0.86 or 0.18 versus 0.84 or 0.20). Other features comparing the two networks are also informative: The graph eccentricity (10.86 vs 10.47), shortest-path length (5.68 vs 5.14), and clustering coefficient (0.29 vs 0.27) all describe the sampled network as more spread out compared with random. On average, the probability sampling algorithm was more efficient than the random sampling algorithm in terms of the number of times workers had to answer each question: $\langle \mathcal{A} \rangle = 2.52$ for the sampling algorithm compared with $\langle \mathcal{A} \rangle = 4.07$ for random. Taken together, the sampling algorithm maintained a comparable level of certainty in the network with far fewer answers on average than the random algorithm.

In general, the sampling algorithm achieved much higher rates of exploration than random while sacrificing only a marginal degree of confidence in question responses.

4.5 Discussion

We study the problem of efficient assignment of crowdsourcing tasks to workers when those workers are able to propose tasks themselves. Using workers to contribute new tasks and not merely perform predetermined tasks helps unlock the true potential of crowdsourcing. We formulate a growing question network model for this problem, prove theoretical properties of this system, and develop and validate Thompson sampling algorithms that can guide workers to grow the network efficiently, while only sacrificing minimal confidence in labels.

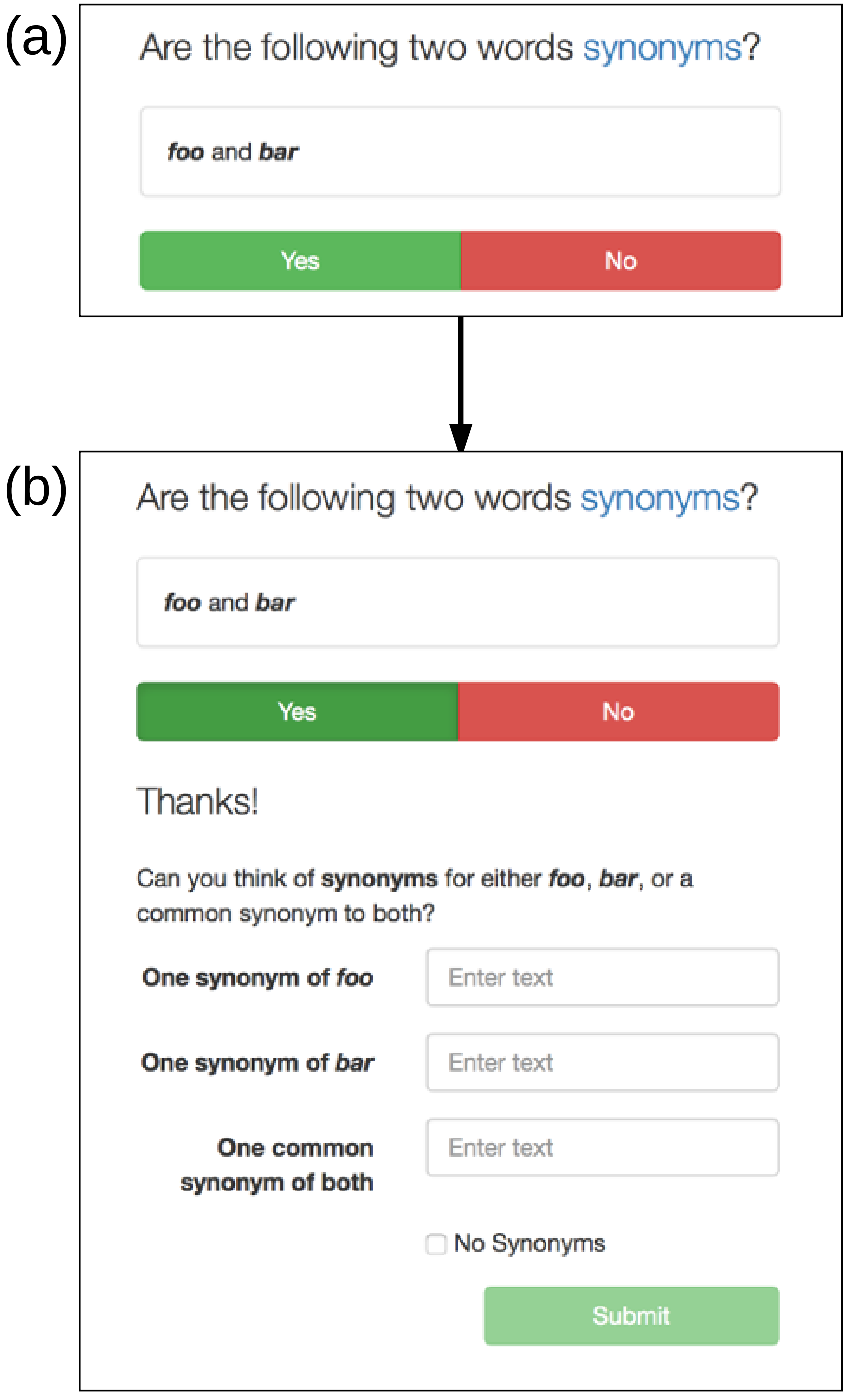
Modeling the evolution of the uncontrolled question network teaches us how to better design crowdsourcing policies. For example, by monitoring the innovation rate (ρ) and exploration rate (η) of the growing question network, a crowdsourcer may be able to better and more efficiently control the question network as it grows. At the same time, the rich-get-richer growth of items (older items are attached to a larger fraction of questions), implies that crowdsourcers should pay special attention to the newest items entering the network, to balance out this bias.

Thompson sampling is fast, easy to implement, and flexible enough to capture the preferences of different crowdsourcers, but it is only one potential policy for question selection. More rigorous question selection techniques can be implemented which may outperform the proposed techniques, but with potentially more restrictions. Further, statistical inference of question parameters and worker features [163], based on extensions of the null model analyzed in Sec. 4.2.2, can be used by the crowdsourcer to better pair workers with questions.

In the future we will address more detailed schemes for question selection. Questions that contain more than a binary (1/0) should be further investigated, although the only adaptation of the above selection scheme is in the choice of metric to Thompson sample from. Different network structures may lend themselves to different problems, and assessing the accuracy of the network inferred by the crowdsourcing will also be investigated.

4.6 Acknowledgments

We thank M. Wagy and V. Rajendran for useful comments and gratefully acknowledge the resources provided by the Vermont Advanced Computing Core. This material is based upon work supported by the National Science Foundation under Grant No. ISS-1447634.



Are the following two words synonyms?

foo and ***bar***

Are the following two words synonyms?

foo and ***bar***

Thanks!

Can you think of **synonyms** for either ***foo***, ***bar***, or a common synonym to both?

One synonym of *foo* Enter text

One synonym of *bar* Enter text

One common synonym of both Enter text

□ No Synonyms

Figure 4.3: Illustration of the Mechanical Turk web interface for the synonymy proposal task.

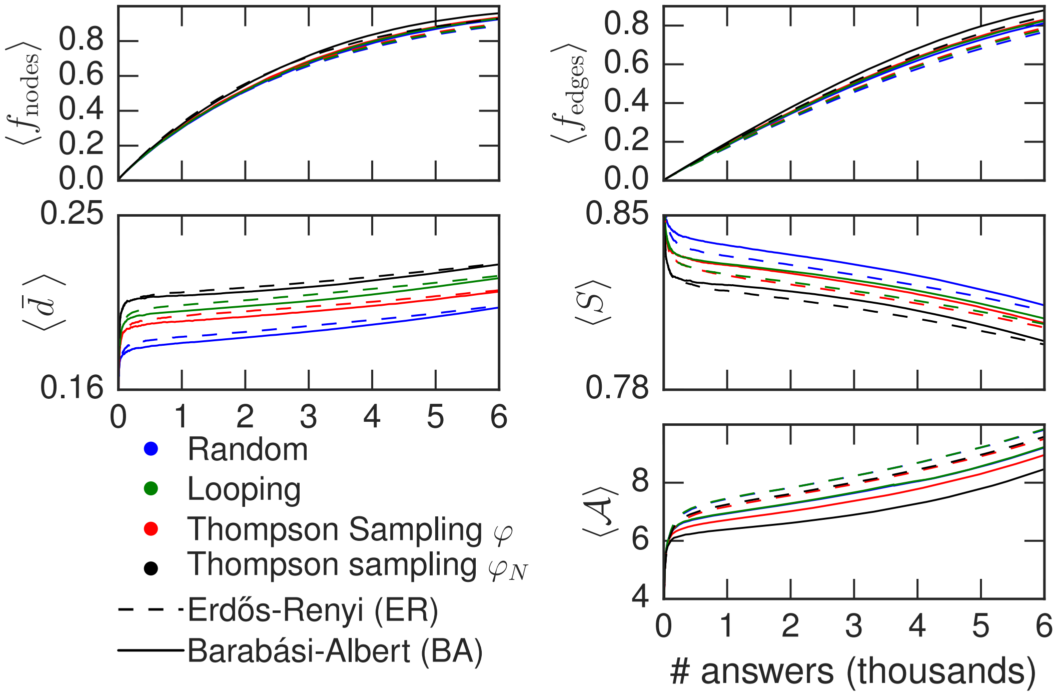
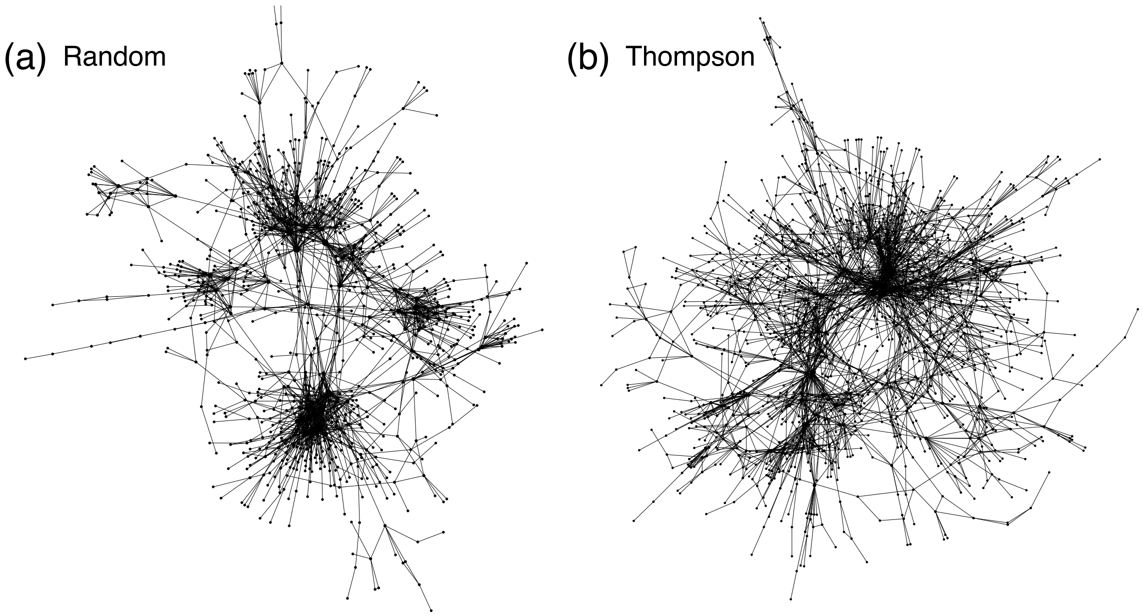


Figure 4.4: Experiment 1's evaluation metrics for four different question selection algorithms.



(c)

	$N(\text{Items})$	$N(\text{Qs})$	$\langle \mathcal{A} \rangle$	$\langle k \rangle$	$\langle CC \rangle$	$\langle e \rangle$	$\langle \ell \rangle$	$\langle S \rangle$	$\langle \bar{d} \rangle$
Random	927	1963	4.07	4.24	0.27	10.47	5.14	0.84	0.20
Thompson Sampling	1464	3103	2.52	4.23	0.29	10.86	5.68	0.86	0.18

Figure 4.5: Comparison of question networks for the synonymy proposal task under random sampling and φ_N sampling.

Chapter 5
Conclusion

In this thesis we show how weighted networks play a crucial role in (i) infrastructure vulnerability, (ii) the study of social phenomena, and (iii) crowd sourcing and control of human labor. In chapter 2, we demonstrate the additional difficulties of constructing robust power grid networks when finite-size effects are important enough to consider. Building on previous work we relate the robustness of a microgrid, weighted by the length of connections, to classical results of infrastructure vulnerability and show that nodes (loads) that carry many connections are also the most vulnerable. Chapter 3 focuses on how our perceptions of causality differ from everyday text on the social network twitter, reemphasizing how the intensity of our connections to one another mold our viewpoints. We find key linguistic factors, including emotional state, differentiate causal language from random conversation and conclude further work in causality must account for emotional bias. Our final look at weighted networks in chapter 4 delved into crowdsourcing and the efficient use of willing volunteers to complete simple tasks. We model a system of objects and ask humans whether or not two object in the system should be linked. After they answer the given question, they are allowed to contribute new potential objects and connection to the system. Finally, this chapter demonstrates sampling techniques are able to efficiently control this evolving, weighted, crowdsourced network.

This thesis establishes three novel projects, tied together using weighted network models and supports future work in micro grid construction, the emotional content of causal statements on a social network, and the efficient control of a group of workers toward solving a hidden set of tasks.

5.1 Weighted network science

Random networks have served as an important tool in discovering similar structures across many scientific disciplines. One of the most important structural characteristics of networks, the scale-free property, shows how many different types of networks all exhibit a similar behavior. Weighted networks build on these guiding principles by considering the intensity of links between two objects, rather than just a simple connection.

In this thesis we show how percolation theory, the study of building and breaking of a system, can suggest important structural characteristics in the construction of microgrid technology. The microgrid could have been modeled without weighted networks, but must less intuitively. A step in a different direction, we find an emotional pattern to causality that could only have been found using the twitter social network. Even though we do not use the explicit intensity of social links between users, all our analysis springs from the underlying heterogeneity of human interaction that Granovetter showed weighted networks easily capture. A capstone of weighted networks and human ingenuity, the weighted network model is responsible for controlling a group of workers to explore an unknown system. Again a more traditional Markov Decision process could model this phenomena, but this type of model would have more awkward formulation compared to the weighted network suggesting a weighted network as a more natural model.

Statistical models of networks and on networks open up flexible ways to analyze a variety of data. Although this thesis used weighted networks as the objects of study, recent advancement in weighted networks focus on how to connect variables in a system to solve more complicated statistical models. A plethora of work in this area has revealed powerful new methods to study causal and highly correlated statistical models.

This work also looked exclusively at weighting the links of a network, but they're are no restrictions on weighting nodes also. Commonly referred to as a node's strength, weighting nodes is normally done by first weighting links and next weighting each node by summing up those link weights that are connected corresponding nodes. Scientists have looked into the statistical distribution of node strength hoping to connect this to specific network functions. Furthermore, setting each link weight to one recovers the traditional node degree and we see then node strength is a generalization.

5.2 Implications of this work

This thesis offers new ways to power our country using weighted network analysis, new insights into our social perceptions of causality, and combines lessons from the two former projects and proposes a new, weighted network, model for the efficient use of an online human workforce. Micro-grid analysis, in chapter 2 further supports the discussion of the traditional utilities role in our energy supply, and how local energy use can take over. Studying our causal perceptions, on a large social network, from a more emotive standpoint offers an alternative view to the classic study of people's causal perceptions of known physical phenomena. Finally, we combine the weighted network analysis with human ingenuity to discover a new way the online workforce can contribute toward scientific goals or complex problems, and set out to prove dynamic weighted network are well-suited for studying a group of workers answering and contributing tasks to a system.

5.2.1 Implications to Power Grid economics

Our work in the powergrid shows that weighted networks resolves difficulties connecting residential houses together into a microgrid, and brings into question the utility companies role in energy use, and whether the traditional utility should restructure the delivery of energy to homes and its business model. The United States powergrid and traditional structuring of the utility connects houses and other loads together in a hierarchical, treelike fashion. This treelike structuring of the powergrid can lead to larger than expected losses of power during catastrophe. This thesis offers a feasible way to construct small distributed microgrids, and future work can build on this model to consider microgrids the next step in power grid restructuring. The microgrid paradigm implies small residential communities use each others power as opposed to using electricity offered from the traditional utility, draining revenue from utility companies and changing the economics of the electricity market.

Micro-grids challenge the traditional utility in three ways: Consumers are in control of their fuel source and often choose cleaner options, economics interests restrict the traditional utilitie's location near fuel sources while the Micro-grid is located in the same neighborhood it serves, and Micro-grid lends itself to smartgrid technologies aimed at minimizing costs (with demand-side response) and maximizing reliability.

Our work further enhances this last advantage of Micro0grids by considering the most reliable placement of powerlines in a small cluster of loads under a budget.

5.2.2 Implications to Causal perception

Second, this thesis explores how sentiment characteristics affect causal attribution and our informal perceptions of causality. To date and to the best of our knowledge, causal perception has been studied from a more objective standpoint, measuring people's written reports to physical phenomena purposely modified by the investigator. For example scientists like Michotte and Peiget attempted to uncover objective measures of how we perceive and understand our world by recording subject's written reports of one ball striking another while modifying the velocities of each object. This work showed people causally attribute the movement of the stationary object to the object, already in motion, that struck it. Although a great deal of work focuses on causal attributions such as Michotte's experiment, researchers shy away from investigating how our emotions skew causal relationships. Our work shows that, as humans, we reserve causal interpretation for strict scientific disciplines like medical research, Natural phenomena in the news, or actions that directly effect us. In addition and unlike experimenting with more objective physical phenomena, our emotional reaction to observing how two objects interact determines whether we are more likely to classify the event as causal or not. These findings imply humans mix causality with emotions, and that our emotional state bends our perceptions of cause and effect relationships. The role our emotions play in causal interpretation and attribution supports a more Humean view of causality over Kant's apriori truth, and adds emotional content to the idea we obtain causal knowledge through covariation. Many argue against Hume's view that we need repeated covariation of two objects to develop causal laws. For example, we quickly learn the causal relationship between pain and touching one's hand on the stove. This first time causal learning, without repeated evidence, begins to break Hume's view of causality. But what about emotion's role. This thesis suggests emotions play a role in causality, and if true, it is possible that the one time experience of burning one's hand on the stove is instead the repeated observation that touching an object causes pain. The emotion pain links the tactile sensation of a group of similar objects (thorns on plants, boiling oil), and appends a burning hot stove to this set. The observation that emotions changes how we view the casual relationship between two objects also implies that emotional content of an observer confounds causal relationships they study, and in future work this thesis recommends (i) experimental testing of how our emotions link causal concepts together and confound the observation that 'first-time' experiences induce causal learning in humans, and (ii) measuring emotional state and further exploring how cause and effect co-vary with changing sentiment.

5.2.3 Implications to Crowdsourcing

This last project combines lessons learned from the previous two projects to introduce a novel model to guide the work flow of a disparate team of workers. Similar to the Micro-grid project, we demonstrate the flexibility of weighted networks to model many different phenomena, and like the causal perceptions project recognize humanities creativity and attempt to harness this power.

The weighted network was a natural choice when modeling a set of objects and their relatedness to one another. As An evolving network of worker proposed and worker validated relationships we easily understand how the system grows by adding nodes (a worker proposing new words) or by adding links (suggesting two already uncovered objects are linked). This ease of interpretation as a network, and the ability to incorporate all the native weighted network metrics allows us to better guide the network as it grows toward under-grown or uncertain parts of the network.

Compared to Markov decision processes, weighted networks better incorporate system growth or the discovery of hidden components in a system. Markov decision processes, by definition, require stating the set of all states and all possible actions a system can take, and this very definition limits the evolution of a system outside the possible occupied states and actions. By contrast, weighted network theory has a long history of evolutionary dynamics and the mathematics of this growth. Expressing crowdsourcing efforts as weighted complex networks, and allowing them to evolve, opens up new paths outside the scope of the Markov decision process.

This project guides human creativity to uncover hidden parts of a system, and implies that in-order to solve a problem each task to complete does not need to be known. We also show how to model this evolutionary process with a weighted network, and find that first-mover effects govern the way a set of humans discover the system. Adding another human element to Crowdsourcing, we bring this field closer to funneling the potential work a human can offer toward the task at hand. When only considering a human's physiological abilities to solve a problem (for instance our ability to easily identify objects inside images), crowdsourcing wastes our other talents. This crowdsourcing analysis and proposal to use weighted networks more efficiently uses each human's time and fosters creativity in the system under study.

Future work in this are can (i) develop new metrics to capture uncertainty and exploration in evolving weighted networks (ii) develop more ways to use weighted networks to use the human potential for guided work on a task, and (iii) explore other human-only attributes to perform more meaningful and efficient work.

5.3 Summary, contribution, and connecting

This thesis studies how networks can guide infrastructure choices, serve as a resource to study social phenomena, and guide a crowd toward efficient exploration and exploitation of a complex human-contributed set of tasks. We use weighted networks to explore the feasibility of residential microgrids, finding nodes with high degree tend to connect to more distant other nodes causing system-wide vulnerabilities. This work in the power grid recommends lowering the effective distance (α) of lines connecting distant nodes, and considering the system robustness gained by creating more distributed, as opposed to the tree-like traditional grid, grid structures. The thesis next shifts toward our perceptions of Causality, and we find that sentiment plays a major role in how we perceive causal links between objects. The last work uses weighted network theory, and lessons learned from studying social networks, to develop a novel way to use human ingenuity and model crowdsourcing. My work shows that the weighted networks serves as a natural object to describe a system of tasks for workers to complete and proposal mechanism for workers to grow the system of tasks effectively. This thesis contributes to our current scientific knowledge on weighted stochastic networks, and their unique application to critical infrastructure, causal perceptions, and crowdsourcing.

Bibliography

[1] Ittai Abraham, Omar Alonso, Vasilis Kandylas, and Aleksandrs Slivkins. Adaptive crowdsourcing algorithms for the bandit survey problem. In *COLT*, pages 882–910, 2013.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.

[3] Alan Agresti and Maria Kateri. *Categorical Data Analysis*. Springer, 2011.

[4] Réka Albert, István Albert, and Gary L Nakarado. Structural vulnerability of the north american power grid. *Phys. Rev. E*, 69(2):025103, 2004.

[5] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[7] Jay Apt, Seth A Blumsack, and Lester B Lave. Competitive energy options for pennsylvania. 2007.

[8] Sergio Arianos, E Bompard, A Carbone, and Fei Xue. Power grid vulnerability: A complex network approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(1):013119, 2009.

[9] J Ash and D Newth. Optimizing complex networks for resilience against cascading failure. *Physica A: Statistical Mechanics and its Applications*, 380:673–683, 2007.

[10] Sitaram Asur and Bernardo A Huberman. Predicting the Future With Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[11] James P Bagrow, Sune Lehmann, and Yong-Yeol Ahn. Robustness and modular structure in networks. *arXiv preprint arXiv:1102.5085*, 2011.

[12] James P Bagrow, Jie Sun, and Daniel ben Avraham. Phase transition in the rich-get-richer mechanism due to finite-size effects. *Journal of Physics A: Mathematical and Theoretical*, 41(18):185001, 2008.

[13] Frank Ball, Denis Mollison, and Gianpaolo Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7(1):46–89, 1997.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[15] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173–187, 1999.

[16] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[17] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[18] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[19] Michael GH Bell and Yasunori Iida. *Transportation network analysis*. 1997.

[20] Kirsten E Bevelander, Kirsikka Kaipainen, Robert Swain, Simone Dohle, Josh C Bongard, Paul DH Hines, and Brian Wansink. Crowdsourcing novel childhood predictors of adult obesity. *PloS one*, 9(2):e87756, 2014.

[21] Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[23] Gerd Bohner, Herbert Bless, Norbert Schwarz, and Fritz Strack. What triggers causal attributions? the impact of valence and subjective probability. *European Journal of Social Psychology*, 18(4):335–345, 1988.

[24] Josh C Bongard, Paul DH Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu. Crowdsourcing predictors of behavioral outcomes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(1):176–185, 2013.

[25] Daren C Brabham. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, 14(1):75–90, 2008.

[26] Roger Brown and Deborah Fish. The psychological causality implicit in language. *Cognition*, 14(3):237–273, 1983.

[27] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[28] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[29] Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

[30] J Mauricio Campuzano, James P Bagrow, and Daniel ben-Avraham. Kleinberg navigation on anisotropic lattices. *Res. Lett. Phys.*, 2008, 2008.

[31] Cécile Caretta Cartozo and Paolo De Los Rios. Extended navigability of small world networks: Exact results and new insights. *Phys. Rev. Lett.*, 102:238703, Jun 2009.

[32] Shai Carmi, Stephen Carter, Jie Sun, and Daniel ben-Avraham. Asymptotic behavior of the kleinberg model. *Phys. Rev. Lett.*, 102(23):238702, 2009.

[33] Benjamin A Carreras, Vickie E Lynch, Ian Dobson, and David E Newman. Critical points and transitions in an electric power transmission model for cascading failure blackouts. *Chaos: An interdisciplinary journal of nonlinear science*, 12(4):985–994, 2002.

[34] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[35] Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *ICML (3)*, pages 64–72, 2013.

[36] Robert M Christley, GL Pinchbeck, RG Bowers, D Clancy, NP French, R Bennett, and J Turner. Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology*, 162(10):1024–1031, 2005.

[37] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical review letters*, 85(21):4626, 2000.

[38] Reuven Cohen, Keren Erez, Daniel ben-Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86(16):3682, 2001.

[39] Eduardo Cotilla-Sanchez, Paul DH Hines, Clayton Barrows, and Seth Blumsack. Comparing the topological and electrical structure of the north american electric power infrastructure. *IEEE Systems Journal*, 6(4):616–626, 2012.

[40] Paolo Crucitti, Vito Latora, and Massimo Marchiori. A topological analysis of the italian electric power grid. *Physica A*, 338(1):92–97, 2004.

[41] Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani. The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924, 2007.

[42] Ian Dobson, Benjamin A Carreras, Vickie E Lynch, and David E Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):026103, 2007.

[43] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PloS one*, 6(12):e26752, 2011.

[44] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268. ACM, 2009.

[45] John C Doyle, David L Alderson, Lun Li, Steven Low, Matthew Roughan, Stanislav Shalunov, Reiko Tanaka, and Walter Willinger. The âĂIJrobust yet fragileâĂİ nature of the internet. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14497–14502, 2005.

[46] Catherine Dwyer, Starr Hiltz, and Katia Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, page 339, 2007.

[47] P Erd0s. Graph theory and probability. *canad. J. Math*, 11:34G38, 1959.

[48] Paul Erd6s and A Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[49] P Erdős. Graph theory and probability. ii. In *Canad. J. Math.* Citeseer, 1960.

[50] Paul Erdős and Alfréd Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[51] Paul Erdős and Alfréd Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1-2):261–267, 1961.

[52] Leonhard Euler. Leonhard euler and the königsberg bridges. *Scientific American*, 189(1):66–70, 1953.

[53] Philip Vos Fellman and Roxana Wright. Modeling terrorist networks, complex systems at the mid-range. *arXiv preprint arXiv:1405.6989*, 2014.

[54] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[55] W Nelson Francis and Henry Kucera. Brown corpus manual. *Brown University*, 1979.

[56] Constantine E Frangakis and Donald B Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002.

[57] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[58] Michael T Gastner and M. E. J. Newman. The spatial structure of networks. *Eur. Phys. J. B*, 49(2):247–252, 2006.

[59] Michael T Gastner and Mark E. J. Newman. Shape and efficiency in spatial distribution networks. *J. Stat. Mech.*, 2006(01):P01015, 2006.

[60] Michael T Gastner and MEJ Newman. Optimal design of spatial distribution networks. *Physical Review E*, 74(1):016117, 2006.

[61] D Gfeller, P De Los Rios, A Caflisch, and F Rao. Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Sciences*, 104(6):1817–1822, 2007.

[62] Roxana Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics, 2003.

[63] Roxana Girju, Dan Moldovan, et al. Text Mining for Causal Relations. In *FLAIRS Conference*, pages 360–364, 2002.

[64] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[65] Clive W J Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[66] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[67] Adolf Grünbaum. Causality and the science of human behavior. *American Scientist*, 40(4):665–689, 1952.

[68] Roger Guimera, Stefano Mossa, Adrian Turtschi, and LA Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

[69] Nikos Hatziargyriou, Hiroshi Asano, Reza Iravani, and Chris Marnay. Microgrids. *Power and Energy Magazine, IEEE*, 5(4):78–94, 2007.

[70] Denis J Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.

[71] Colin Hines and David Lowry. Communication on energy: Nuclear power and the non proliferation treaty: how the carrot became a turnip. *Energy Policy*, 13(6):592–593, 1985.

[72] Paul Hines, Jay Apt, and Sarosh Talukdar. Large blackouts in north america: Historical trends and policy implications. *Energy Policy*, 37(12):5249–5259, 2009.

[73] Rose Holley. Crowdsourcing: how and why should libraries do it? *D-Lib Magazine*, 16(3):4, 2010.

[74] Joseph Hoshen and Raoul Kopelman. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Phys. Rev. B*, 14(8):3438, 1976.

[75] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[76] Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 94(12):128701, 2005.

[77] David Hume. *A Treatise of Human Nature*. Courier Corporation, 2012.

[78] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering–WISE 2013*, pages 1–15. Springer, 2013.

[79] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. Quizz: targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on World wide web*, pages 143–154. ACM, 2014.

[80] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[81] RB Joynson. Michotte's Experimental Methods. *British Journal of Psychology*, 62(3):293–302, 1971.

[82] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[83] Immanuel Kant and Paul Guyer. *Critique of Pure Reason*. Cambridge University Press, 1998.

[84] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of Social Media. *Business horizons*, 53(1):59–68, 2010.

[85] David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[86] Ehud D Karnin, Eugene Walach, and Tal Drory. *Crowdsourcing in the document processing practice*. Springer, 2010.

[87] F Katiraei and MR Iravani. Power management strategies for a microgrid with multiple distributed generation units. *Power Systems, IEEE Transactions on*, 21(4):1821–1831, 2006.

[88] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

[89] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[90] Harold H Kelley. Attribution Theory in Social Psychology. In *Nebraska symposium on motivation*. University of Nebraska Press, 1967.

[91] Harold H Kelley and John L Michela. Attribution Theory and Research. *Annual review of psychology*, 31(1):457–501, 1980.

[92] Faiza Khan Khattak and Ansaf Salleb-Aouissi. Quality control of crowd labeling through expert evaluation. In *Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds*, 2011.

[93] Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 885–890. ACM, 2013.

[94] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[95] Dan Klein and Christopher D Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, volume 15, page 3. MIT Press, 2003.

[96] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[97] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, 7(1):48–50, 1956.

[98] Heidi J Larson, Louis Z Cooper, Juhani Eskola, Samuel L Katz, and Scott Ratzan. Addressing the vaccine confidence gap. *The Lancet*, 378(9790):526–535, 2011.

[99] Robert H Lasseter and Paolo Paigi. Microgrid: a conceptual solution. In *Power Electronics Specialists Conference, 2004. PESC 04. 2004 IEEE 35th Annual*, volume 6, pages 4285–4290. IEEE, 2004.

[100] Vito Latora and Massimo Marchiori. How the science of complex networks can help developing strategies against terrorism. *Chaos, solitons & fractals*, 20(1):69–75, 2004.

[101] David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[102] G. Li, S. D. S. Reis, A. A. Moreira, S. Havlin, H. E. Stanley, and J. S. Andrade. Towards design principles for optimal transport networks. *Phys. Rev. Lett.*, 104:018701, Jan 2010.

[103] Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J Quinn. Crowdsourcing high quality labels with a tight budget. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 237–246. ACM, 2016.

[104] Nan Lin and Mary Dumin. Access to occupations through social ties. *Social networks*, 8(4):365–385, 1986.

[105] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.

[106] Julie B Lovins. *Development of a Stemming Algorithm.* MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[107] R Frederick Ludlow and Sijbren Otto. Systems chemistry. *Chemical Society Reviews*, 37(1):101–108, 2008.

[108] Diana Lynn MacLean and Jeffrey Heer. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, 20(6):1120–1127, 2013.

[109] P Achintya Madduri, Jason Poon, Javier Rosa, Matthew Podolsky, Eric Brewer, and Seth Sanders. A scalable dc microgrid architecture for rural electrification in emerging regions. In *2015 IEEE Applied Power Electronics Conference and Exposition (APEC)*, pages 703–708. IEEE, 2015.

[110] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[111] Sonja Marjanovic, Caroline Fry, and Joanna Chataway. Crowdsourcing based business models: In search of evidence for innovation 2.0. *Science and public policy*, page scs009, 2012.

[112] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[113] Andrew Kachites McCallum. MALLETT: A Machine Learning for Language Toolkit. 2002.

[114] Doug Mckenzie-Mohr. New ways to promote proenvironmental behavior: Promoting sustainable behavior: An introduction to community-based social marketing. *Journal of social issues*, 56(3):543–554, 2000.

[115] Lauren Ancel Meyers, MEJ Newman, Michael Martin, and Stephanie Schrag. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. *Emerging infectious diseases*, 9(2):204–210, 2003.

[116] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[117] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[118] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PloS one*, 8(5):e64417, 2013.

[119] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algor.*, 6(2-3):161–180, 1995.

[120] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[121] M. E. J. Newman. *Networks: an introduction.* Oxford University Press, 2010.

[122] M. E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[123] M. E. J. Newman and Duncan J Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332, 1999.

[124] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[125] Jonathan R Nitschke. Systems chemistry: Molecular networks come of age. *Nature*, 462(7274):736–738, 2009.

[126] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[127] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701, Nov 2001.

[128] Judea Pearl. *Causality.* Cambridge university press, 2009.

[129] Chaveevan Pechsiri and Asanee Kawtrakul. Mining Causality from Texts for Question Answering System. *IEICE TRANSACTIONS on Information and Systems*, 90(10):1523–1533, 2007.

[130] Steven Pinker. *The Better Angels of Our Nature: Why Violence Has Declined.* Penguin, 2011.

[131] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A Rule based Approach to Word Lemmatization. *Proceedings of IS-2004*, pages 83–86, 2004.

[132] Wouter Poortinga, Linda Steg, and Charles Vlek. Values, environmental concern, and environmental behavior a study into household energy use. *Environment and behavior*, 36(1):70–93, 2004.

[133] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

[134] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[135] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM, 2012.

[136] Vaibhav Rajan, Sakyajit Bhattacharya, L Elisa Celis, Deepthi Chander, Koustuv Dasgupta, and Saraschandra Karanam. Crowdcontrol: An online learning approach for optimal task scheduling in a dynamic crowd platform. In *Proceedings of ICML Workshop: Machine Learning Meets Crowdsourcing*, 2013.

[137] Benjamin L Ranard, Yoonhee P Ha, Zachary F Meisel, David A Asch, Shawndra S Hill, Lance B Becker, Anne K Seymour, and Raina M Merchant. Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review. *Journal of general internal medicine*, 29(1):187–203, 2014.

[138] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*, 2011.

[139] Mickey R Roberson and Daniel ben-Avraham. Kleinberg navigation in fractal small-world networks. *Phys. Rev. E*, 74(1):017101, 2006.

[140] Martin Rolfs, Michael Dambacher, and Patrick Cavanagh. Visual Adaptation of the Perception of Causality. *Current Biology*, 23(3):250–254, 2013.

[141] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM, 2010.

[142] Alejandro F. Rozenfeld, Reuven Cohen, Daniel ben Avraham, and Shlomo Havlin. Scale-free networks on lattices. *Phys. Rev. Lett.*, 89:218701, Nov 2002.

[143] Donald B Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 2011.

[144] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[145] Richard M Ryan and James P Connell. Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of personality and social psychology*, 57(5):749, 1989.

[146] Marcel Salathé and Shashank Khandelwal. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol*, 7(10):e1002199, 2011.

[147] Christian M Schneider, André A Moreira, José S Andrade, Shlomo Havlin, and Hans J Herrmann. Mitigation of malicious attacks on networks. *Proc. Natl. Acad. Sci. USA*, 108(10):3838–3841, 2011.

[148] Brian J Scholl and Patrice D Tremoulet. Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8):299–309, 2000.

[149] Jasjeet S Sekhon. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods. *The Oxford handbook of political methodology*, pages 271–299, 2008.

[150] Louis M Shekhtman, James P Bagrow, and Dirk Brockmann. Robustness of skeletons and salient features in networks. *Journal of Complex Networks*, 2(2):110–120, 2014.

[151] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[152] Richard M Shiffrin. Drawing causal inference from big data. *Proceedings of the National Academy of Sciences*, 113(27):7308–7309, 2016.

[153] Mark DF Shirley and Steve P Rushton. The impacts of network topology on disease spread. *Ecological Complexity*, 2(3):287–299, 2005.

[154] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[155] C Smallwood. Distributed generation in autonomous and nonautonomous micro grids. In *Rural Electric Power Conference, 2002. 2002 IEEE*, pages D1–D1_6. IEEE, 2002.

[156] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[157] Dietrich Stauffer and Amnon Aharony. *Introduction to percolation theory.* Taylor and Francis, 1991.

[158] Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002.

[159] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[160] David MJ Tax, Martijn van Breukelen, Robert P W Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern recognition*, 33(9):1475–1485, 2000.

[161] Shelley E Taylor and Susan T Fiske. Point of View and Perceptions of Causality. *Journal of Personality and Social Psychology*, 32(3):439, 1975.

[162] Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P Bebber, Mark D Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327(5964):439–442, 2010.

[163] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. Efficient crowdsourcing of unknown experts using multi-armed bandits. In *European Conference on Artificial Intelligence*, pages 768–773, 2012.

[164] Long Tran-Thanh, Matteo Venanzi, Alex Rogers, and Nicholas R Jennings. Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 901–908. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[165] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[166] Jiaoe Wang, Huihui Mo, Fahui Wang, and Fengjun Jin. Exploring the network structure and nodal centrality of chinaâĂŹs air transport network: A complex network approach. *Journal of Transport Geography*, 19(4):712–721, 2011.

[167] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[168] Duncan J Watts. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. USA*, 99(9):5766–5771, 2002.

[169] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440–442, 1998.

[170] Eric W Weisstein. Square line picking. *From MathWorld–A Wolfram Web Resource*, 2005. mathworld.wolfram.com/SquareLinePicking.html.

[171] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[172] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who Says What to Whom on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011.

[173] Yongxiang Xia, Jin Fan, and David Hill. Cascading failure in watts–strogatz small-world networks. *Physica A: Statistical Mechanics and its Applications*, 389(6):1281–1285, 2010.

[174] Zengwang Xu and Robert Harriss. Exploring the structure of the us intercity passenger air transportation network: a weighted complex network approach. *GeoJournal*, 73(2):87–102, 2008.