2021

# Exploring Hidden Networks Yields Important Insights in Disparate Fields of Study

Laurence Clarfeld
*University of Vermont*

# Exploring Hidden Networks Yields Important Insights in Disparate Fields of Study

A Dissertation Presented

by

Laurence A. Clarfeld

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 Computer Science

January, 2021

Defense Date: November 9th, 2020
Dissertation Examination Committee:

Margaret J. Eppstein, Ph.D., Advisor
Robert Gramling, MD, DSc, Chairperson
Paul D.H. Hines, Ph.D.
Donna M. Rizzo, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

# Abstract

Network science captures a broad range of problems related to things (nodes) and relationships between them (edges). This dissertation explores real-world network problems in disparate domain applications where exploring less obvious "hidden networks" reveals important dynamics of the original network.

The power grid is an explicit network of buses (e.g., generators) connected by branches (e.g., transmission lines). In rare cases, if $k$ branches (a $k$-set) fail simultaneously, a cascading blackout may ensue; we refer to such $k$-sets as "defective". We calculate system risk of cascading failure due to defective 2-sets and 3-sets in synthetic test cases of the Polish and Western US power grids. A stochastic group testing algorithm (Random Chemistry) is used to efficiently sample defective $k$-sets in the "hidden network" of all possible $k$-sets, and new methods are proposed to derive bounds on the total number of defective sets from the obtained sample. We use copula analysis, with a custom distance metric, to estimate risk when the $k$ initiating outages are spatially correlated and show that this realistic assumption increases the relative contribution to risk of 3-sets over 2-sets.

In the power systems application, among others, computational costs vary when testing defective *vs.* non-defective $k$-sets, a consideration that has not previously been made when evaluating group testing algorithms. We develop a domain-independent test problem generator that enables us to vary the number of defective $k$-sets, with a tunable parameter to control the cost ratio of testing defective *vs.* non-defective $k$-sets. We introduce a deterministic group-testing algorithm (SIGHT) capable of sampling from this space, and show that both the number of defective sets and the test cost ratio affect the relative efficiency of Random Chemistry *vs.* SIGHT. We discuss various applications where each algorithm is expected to outperform the other.

Conversations can also be viewed as explicit networks of dialog (edges) between speakers (nodes). We propose using second and third order Markov models based on the sequence of speaker turn lengths to elucidate "hidden networks" of information flow and reveal patterns of information sharing between participants. The proposed method is demonstrated on a corpus of conversations between patients with advanced cancer and palliative care clinicians. We demonstrate the efficacy of the model by confirming known patterns of conversational discourse, identifying normative patterns of information flow in serious illness conversation, and showing how these patterns differ under a variety of contexts, including the expression of distressing emotion.

# Citations

Material from this dissertation has been published in the following form:

### Chapter 2

Clarfeld, L. A., Hines, P. D. H., Hernandez, E. M. and Eppstein, M. J.. "Risk of Cascading Blackouts Given Correlated Component Outages," in *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1133-1144, 1 July-Sept. 2020, doi: 10.1109/TNSE.2019.2910837.

Clarfeld, L. A., Eppstein, M. J., Hines, P. D. H. and Hernandez, E. M.. "Assessing Risk from Cascading Blackouts Given Correlated Component Failures," *2018 Power Systems Computation Conference (PSCC)*, Dublin, 2018, pp. 1-7, doi: 10.23919/PSCC.2018.8442655.

### Chapter 3

Clarfeld, L. A., and Eppstein, M. J.. "Group Testing for Efficiently Sampling Hypergraphs When Tests Have Variable Costs," in invited revision for *IEEE Transactions on Network Science and Engineering*, reviews received November 1, 2020. (Preprint available at arXiv:2010.09205. 2020 Oct 11.)

### Chapter 4

Clarfeld, L. A., Gramling, R., Rizzo, D. M., and Eppstein, M. J.. "A General Model of Conversational Dynamics and an Example Application in Serious Illness Communication," in review at *PLOS One*, submitted September 27, 2020. (Preprint available at arXiv:2010.05164. 2020 Oct 11.)

### Chapter 5

Clarfeld, L. A., and Eppstein, M. J.. Group-Testing on Hypergraphs with Variable-Cost Tests: A Power Systems Case Study. arXiv:1909.04513. 2019 Sep 9.

# Acknowledgements

I would first like to thank my advisor, Maggie Eppstein, for mentoring me, guiding me through my doctoral studies, and building my confidence as a researcher. The lessons you've taught me are too numerous to recall. And to the rest of my defense committee, thanks to Paul Hines for co-advising me during my first two years and providing support and deep insight into the power systems aspects of my dissertation. Bob Gramling, thank you for welcoming me into the Vermont Conversation Lab, for our engaging conversations about conversations, and for creating an environment for my research to flourish. Thanks to Donna Rizzo for your support, feedback, and positive energy that helped keep me motivated.

I owe thanks to my other collaborators and colleagues, including Eric Hernandez for his expertise in risk analysis. Thanks to Jeff Buzas and Damin Zhu for their inspiration and engaging discussions about group testing. Thanks to Stewart Alexander and Lukas Ingersoll from Purdue University for providing the data on distressed emotions, which was pivotal to my research. Thanks to the RA's of the Vermont Conversation Lab, Cailin Gramling, Laura Hirsch, Jack Straton, and Brigitte Durieux for their assistance preparing the PCCRI transcripts for analsyis, and to all the other VCL members for their helpful feedback. Thanks to Andy Klem for engaging discussions on cascading failures and set size estimation of defective sets.

I would also like to thank those who supported me outside of academia. Thanks to my parents for always being there for me and to Natalie for always having my back. Lexi and Ezra, you keep me going!

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Comprehensive Introduction

Networks represent relationships between things. It is no surprise that such a simple concept is ubiquitous in nature. While networks have been around since time immemorial, the great mathematician Leonhard Euler is widely credited with formalizing the study of networks (graph theory) in 1735 [1]. Euler was challenged with finding a walking path that crossed each of seven bridges in Königsberg (near modern-day St. Petersburg) exactly once, and in formulating a solution, abstracted the problem by considering the things (land masses) and relationships between them (whether two land masses are connected by a bridge) as a mathematical object.

Since Euler's time, network science has matured as an important field of study with profound impacts on the many disciplines where it is applied. The multidisciplinary nature of the field has been embraced by the network science community [2, 3]. Some popular application domains include biology, social sciences, computer science, engineering, and linguistics (Table 1.1).

Formally, a network (or graph) consists of a set of vertices (nodes) and edges that connect pairs of vertices. Both nodes and edges can be assigned weights that

| Discipline | Applications |
|---|---|
| biology | metabolic networks [4, 5], genetic regulatory networks [6], food webs [7, 8], pollinator networks [9, 10] |
| social science | social networks [11, 12, 13] |
| computer science | wireless networks [14], the internet [15, 16] |
| engineering | infrastructure networks (transportation [17], electricity [18], communication [19]) |
| linguistics | Language networks [20, 21] |

*Table 1.1: Multidisciplinary examples of various applications of network science.*

indicate the strength of the relationships they represent. In the preceding definition, an edge connects two (not necessarily distinct) vertices, but the graph concept can be generalized. In a hypergraph, each edge can connect an arbitrary number of vertices.

In some cases, networks are readily apparent. Their nodes may represent physical objects, where the relationships between them can be directly observed. Over the past several decades, increases in computational power and improved computational methods have allowed for the analysis of larger and more complex networks than at any previous point. For certain problems, however, the outward-facing network is insufficient by itself to formulate a solution. In some cases, less obvious "hidden networks" may be defined to capture the dynamics of the original network. In this dissertation we develop original methods to tackle problems where "hidden networks" are used to gain insights about the original networks, with an emphasis on problems from two specific application domains: (1) How to quantify the risk to the power grid from cascading failures; and (2) How to model information flow in conversations.

Application-specific context is included in subsequent chapters, so in the remainder of this Chapter, we introduce the problems which motivate this dissertation and explain how "hidden networks" can be exploited. We then provide additional background on

methodologies used in subsequent chapters, beyond what is already discussed in the chapters themselves.

## 1.1 Power Networks

The electrification of society is regarded as one of the greatest engineering achievements of the last century, with the US power grid considered the largest interconnected machine in the world [22]. The explicit network structure of the power grid is comprised of generators, loads, and substations (nodes, a.k.a. buses) connected by transmission lines and transformers (edges, a.k.a. branches, or lines). Electricity is extracted from and inserted into the network at buses and transported between buses by lines. The physics of the grid necessitate that the instantaneous amount of power being produced be equal to the amount being consumed at any given moment.

When a set of $k$ out of $N$ lines fail together, the electricity that was flowing through those lines must be re-routed, which may cause additional failures when other lines exceed their operating capacities. Within the power systems literature, such sets are sometimes referred to as $N-k$ contingencies. We also refer to these using the more general term $k$-sets. In rare cases, even the failure of small $k$-sets may result in a cascade that can propagate through the network resulting in a large blackout. We refer to such $k$-sets as $N-k$ malignancies, or by the more general term "defective sets". Due to their vast size and substantial social and economic costs, the risk that large cascading failures pose to power systems is significant [23, 24, 25].

Quantifying the risk posed by defective $k$-sets is challenging, since the number of $N-k$ contingencies grows exponentially with the size of the network. The size

of the search space and the relative scarcity of malignancies makes it infeasible to efficiently identify them all. Thus, an efficient sampling strategy is required. From a sample of defective sets, it is necessary to estimate the total number of defective sets (at least up to a given maximum $k$), in order to estimate system risk. When the failures of individual components of a defective set exhibit correlations, as they are likely to when they share a common cause (such as a storm), one should account for this in estimating risk. These problems are introduced in the following subsections.

### 1.1.1 Group testing and the "hidden" network

In order to efficiently sample defective sets (to estimate risk), the 'Random Chemistry' algorithm has been previously demonstrated to be effective, and so is used in Chapter 2. While not formally recognized as such when it was initially proposed, Random Chemistry is a stochastic, adaptive group-testing algorithm capable of sampling defective hyperedges from the "hidden network" of all $k$-sets.

The main premise behind group testing is that rather than test each element (or subset) individually to determine which are defective, strategies are developed to more rapidly find defective sets by testing pooled samples where it can be determined whether the sample contains at least one defective subset. More formally, given the universal set of elements $V = \{v_1, \ldots, v_N\}$, consider the set $D$ of all minimal defective subsets of $V$, where minimal means that no smaller subset is also defective. Then, for any set $S \subseteq V$ there exists an oracle function $TEST$ such that:

$$TEST(S) = \begin{cases} 1 & \text{if } S \cap D \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

Using $TEST$, the classic group testing research has focused on developing algorithms for identifying defective elements and minimal defective subsets using the fewest number of tests possible, by applying deterministic methods to iteratively test pooled samples of $V$. The process continues until one (or all) minimal defective set(s) are found, depending on the application. Deterministic group testing has been applied in a variety of applications areas, including multichannel access [26, 27], DNA library testing and sequencing [28], and diagnosis of infectious diseases such as influenza [29], chlamydia and gonorrhea [30], HIV [31], and Covid-19 [32, 33]. In adaptive group testing, the results of each test are considered when determining which subsequent tests to perform. The ability to adapt the sequence of tests in response to previous test results obtains greater efficiency than a fixed strategy, but there are some applications where resources are limited and the responsiveness required for adaptive group testing may be infeasible. In these cases, non-adaptive group testing strategies where the sequence (and number) of tests is predetermined may still offer gains in efficiency over testing each set individually.

While the common goal in group testing has been to minimize the total number of tests required, this does not take into account that there may be different costs associated with testing defective *vs.* non-defective sets (as is true for the power systems problem). Furthermore, deterministic group-testing algorithms have not been developed to handle problems with intractably large numbers of defective sets

with variable $k$, where false-negative tests occur and an upper-bound on the size of returned defective sets can be specified (all of which are true for the power systems problem). Thus, it was not clear whether a deterministic, adaptive group testing algorithm might perform better than the stochastic, adaptive group testing algorithm Random Chemistry, in sampling from the hypergraph of all $k$-sets in an application such as the power systems problem. In Chapter 3, we modify a current state-of-the-art deterministic group testing algorithm to handle these issues and develop a tunable, domain-independent, test-problem generator that is used in a first-of-its-kind comparison between a stochastic and a deterministic group testing strategy with variable test costs.

### 1.1.2 Estimating Defective Set Counts

The number of defective $k$-sets in the power systems application is too large for us to be able to identify them all, even in moderately sized power grids. Thus, when using a group testing approach to efficiently sample for defective $k$-sets, we can only calculate an estimate of the total risk to the grid by first estimating the total number of defective $k$-sets that exist, up to some maximum $k$, based on the collected sample of them. (We note that Monte Carlo sampling methods do not require this, but are much less efficient at finding defective $k$-sets than the group-testing approach [34, 35].)

The use of random sampling for predicting total set size has been widely studied in a variety of contexts. In the biological sciences, mark-recapture describes a set of methods that use the rate at which individuals are re-sampled from a population to derive an estimate of population size [36]. The same methods can also be used to estimate species richness [36]. Parametric methods of estimation (such as the non-

linear curve fitting method used in previous work [35]) have been shown ineffective in most practical situations where the underlying distribution of sampling probabilities is unknown [37, 38, 39]. We have observed that some defective sets are more likely to be identified *via* group testing than others, and so non-parametric methods are preferred.

Statistically, most of the information about undetected individuals is contained in the frequencies of the least-encountered individuals in the sample, and thus most of the popular methods use these frequencies to derive population estimates [39]. Because some individuals in the population may have such low probabilities of capture as to be undetectable, these estimators can only be expected to achieve a lower-bound on the true number of individuals in a population [39]. In preliminary work, we experimented with several methods, including jackknifing [40, 41], Chao's improved method [38], and coverage-estimator based methods [42, 43]; ultimately, we found Chao's method [44] to be most effective as a means of lower-bounding the estimate of the number of $N - 3$ malignancies in the power systems problem, so we use this in Chapter 2. To derive an upper-bound on this estimate, a novel method is proposed based on patterns of sampling bias in group testing algorithms observed when the frequency distribution of components within defective sets is heavy-tailed.

### 1.1.3 Risk for Correlated Component Outages

Given a sample of defective sets, and an estimated total count of defective sets, we propose a copula method for estimating total risk in Chapter 2 for correlated component outages. This method has two user-defined parts: a correlation function that determines the strength of correlation between two components and a copula function that determines how the probabilities of failure for each component of a

defective set, along with the pairwise correlation between components of the set, are combined to derive a joint probability of the whole set failing together.

The fundamental assumption in the copula method is that components that are 'closer together' are more likely to fail together than components that are 'farther apart'. In the case of common-cause events such as storms, spatially correlated damage is expected [45], and this is what we assume for the proof-of-concept demonstration in Chapter 2. However, the notion of distance should be interpreted more broadly. For example, the distance apart in time of manufacture or installation for two components can also be thought of as a form of (temporal) distance.

Euclidean distance is a measure of the shortest straight-line distance between two points, making it a natural and intuitive way to measure the spatial distance between two objects. Euclidean distance may capture the distance between two components of the grid "as the crow flies," but neglects topological and electrical considerations. When we abstract the power grid to a graph, we can consider the shortest path between two nodes (or edges) as the quickest way to navigate along the edges of the network from one component to the other. There are several methods, such as Djikstra's algorithm, for efficiently calculating this distance for weighted or unweighted graphs. Measures of electrical distance have been proposed to bridge the gap between the topological structure of the network and the electrical properties of the grid [46] and may better capture how changes to one component of the grid may affect another.

However, when modeling spatially-correlated damage caused by a common exogenous source, such as a storm or earthquake, spatial distance is more relevant than electrical distance. When measuring distance between line segments, rather than points (as we desire to do for quantifying the distance between two transmission lines),

it is unclear how Euclidean distance can be applied. In Chapter 2, we propose a semi-metric such that the more closely the span of two lines overlap, the closer the distance between them.

The correlation function controls how the strength of correlation varies with distance. The correct correlation function will depend on the exogenous, common-cause event being modeled. If, for example, the risk associated with the failure of some software component of the grid is being modeled, correlation may be captured by a Poisson-type distribution [47]. We demonstrate our method assuming spatial correlation modeled with a negative exponential distribution, as further justified in Chapter 2.

Given the pairwise correlation of each branch pair, the copula function 'couples' the probabilities of failure for each branch. Many classes of copula functions exist and which is most appropriate for a particular application depends on the underlying distribution of failure likelihood for each component (eg., [48, 49, 50]). The use of copula functions was popularized in finance [51], and today they are applied in a wide variety of contexts such as hydrology and water resource engineering [52], multi-source image segmentation [53], and precipitation modeling [54, 55]. We demonstrate our method in Chapter 2 as a proof-of-concept, assuming the familiar Gaussian Copula function.

## 1.2 Conversation Networks

In contrast to the power grid, whose dynamics are governed by the laws of physics, a conversation might seem disorderly and unpredictable. On the contrary, conversations

consist of sequences of highly coordinated turns, as described by sociologist Harvey Sacks [56]. Our goal in this application domain is to develop a tool for understanding and interpreting information flow in a conversation, with immediate applications to analyzing conversations between seriously ill patients and palliative care clinicians.

A conversation can be abstracted as an explicit network by considering each speaker as a node and interactions between speakers as edges. This explicit network has been used as a tool for visualizing the strength of relationships between speakers for pre-analysis of conversational dynamics [57]. Given the importance of sequence to understanding conversation, methods for further analysis of conversational dynamics must be capable of measuring how the conversation changes over time. Markov models are one way to capture this sequential dependence relationship between turns.

### 1.2.1 Markov Models and the "Hidden Network"

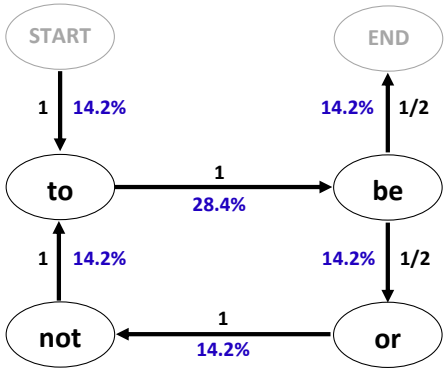
Simply put, Markov models are used to model how a system probabilistically transitions between a finite number of possible states. For example, given the phrase "to be or not to be", we can define a Markov model where each word represents a unique state, and the probability of transitioning from one state to another is determined from the observed word sequence (Fig. 1.1a). The state space diagram is typically labeled such that the sum of the weights of all outgoing edges from each node equals one (as shown by the black edge weights in Fig. 1.1). Such weights can be used as probabilities to generate new sequences of words, or to determine if a given word sequence could have been generated with the model (e.g., the sequence "to be or not to be or not to be" could have been generated using the probabilities shown in Fig. 1.1a).

For our application, however, we "populate" the edges of Markov models with

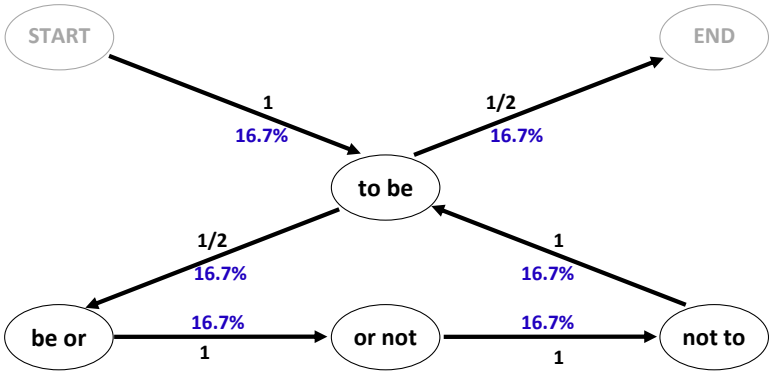
weights that represent percentages with which each transition is observed in some data set, such that the sum of all edges equals 100% (as shown by the blue edge weights in Fig. 1.1). These weights thus quantify patterns of occurrence in the observed data (e.g., in this example, they quantify the transitions in the word sequence "to be or not to be" but not the word sequence "to be or not to be or not to be").

By definition, the current state of a Markov model is dependent only on the previous state, however higher-order models can be considered by simply redefining the state of the system based on the previous $k$ states (called the $k^{\text{th}}$-order model). Fig. 1.1b shows a $2^{\text{nd}}$-order model of the same word sequence "to be or not to be", where each state is defined by a sequence of two words (only those two-word states that are observed in the given word sequence are shown in the figure). As the order of a model increases, the number of possible states (and transitions) grows exponentially, so there are practical limitations to how much memory a model can incorporate. Various methods exist for determining the optimal model order given a particular set of observations [58, 59, 60, 61]. In some cases, higher dimensionality of some states may prove useful while others will not. In these cases, variable-order Markov models may be appropriate [62].

Hidden Markov Models are special types of Markov model, where the model is used to predict some unknown (hidden) state(s). Hidden Markov models have been applied to a wide variety of prediction and classification tasks such as pattern recognition [63, 64, 65], protein structure prediction [66, 67, 68], speech emotion recognition [69], and part-of-speech tagging [70, 71]. Hidden Markov models have also been applied to prediction/classification problems in conversation analysis in a variety of ways (see Chapter 4 for examples).




*(a) 1st-order Markov Model*




*(b) 2nd-order Markov Model*

*Figure 1.1: State space diagrams for a simple example of (a) 1st- and (b) 2nd-order Markov models based on the observed word sequence "to be or not to be." Edge weights show the probability of selecting an outgoing edge from a given node (shown in black) or the overall percentage occurrence for each transition (shown in blue), for this individual phrase. Transitions that did not occur in the observed sequence are not shown. Although there are 16 possible 2nd-order states using the 4 words "to", "be", "or", "not", states with 0% occurrence in this phrase (e.g., "to or", and "be to") are not shown here.*

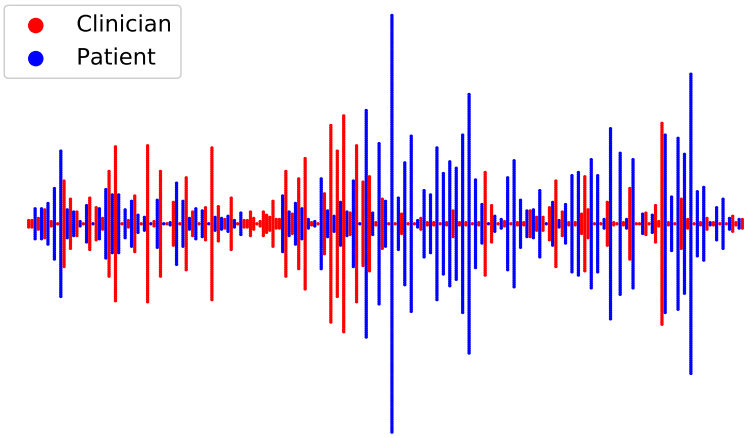
Despite sharing the word "hidden", the "hidden network" that we will consider to help understand information flow should not be confused with hidden Markov models. Our goal is interpretation, rather than prediction or classification. Specifically, we model a "hidden network" of information sharing patterns using a Markov model based on binarized turn lengths, with turn length being a proxy for the amount of information exchanged each speaker turn.

Figure 1.2 shows a context-free, graphical representation of (a) a real conversation between a palliative care clinician and patient with advanced cancer and (b) a fictitious conversation with two speakers and twelve speaker turns. The real conversation shows naturally occurring patterns, such as the oscillation between short and long turns, that we seek to capture with our model. Given the threshold of eight words, the fictitious conversation is used to show how each turn can be binarized such that turns with fewer than eight words are short (S) and turns with eight or more words are long (L). In our example, this results in the turn-length sequence {S, S, S, L, S, L, S, L, L, L, S, S}.

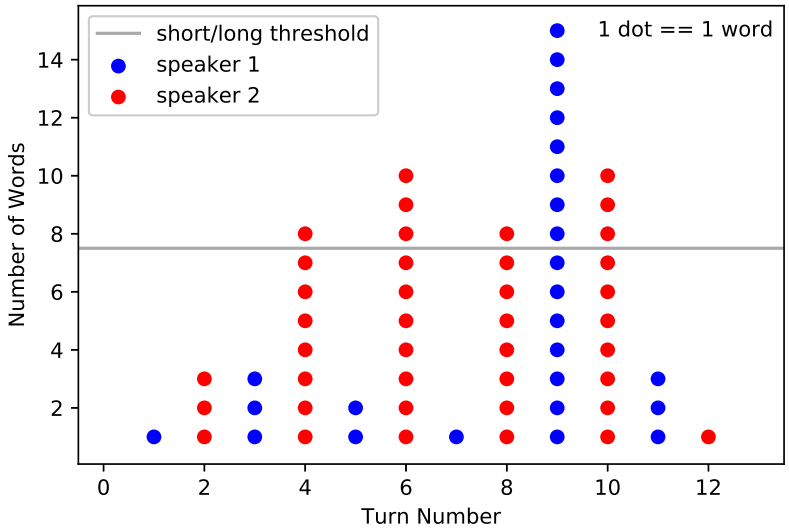
In Chapter 4, we introduce COnversational DYnamics Models (CODYMs), a novel application of Markov models populated using observed sequences of short/long turns. We use 2$^{nd}$- and 3$^{rd}$-order CODYMs to analyze information flow in conversations between patients with advanced cancer and palliative care clinicians.

## 1.3 Dissertation Outline

This dissertation adopts a 'journal format' and thus the next three chapters represent stand-alone, distinct works, each prepared individually for publication. Chapter




*(a) A graphical representation of an actual conversation between a clinician and a patient with advanced cancer, where each vertical column represents a speaker turn, the height of the column is proportionate to the number of words in that turn, and the color of each column indicates the speaker (clinician-side turns in red, patient-side turns in blue).*




*(b) A graphical representation of a conversation, where each dot represents a word, each vertical column of dots represents a speaker turn (colored by speaker), and a pre-determined threshold (here, eight words) can be used to classify turns as short (if the turn contains fewer words than the threshold) or long.*

*Figure 1.2*

2 presents the copula method for determining the risk of cascading failures when component outages are correlated. The method is demonstrated on two test cases, including a synthetic model of the Western US power grid. To apply the method to this large test case, new methods were developed for estimating lower and upper bounds on the number of defective 3-sets from a sample of such sets obtained *via* stochastic, adaptive group testing. Chapter 3 gives further consideration to group testing algorithms that can be used to identify defective hyperedges in hypergraphs, demonstrating that the cost of defective *vs.* non-defective tests, and the prevalence of defective $k$-sets, determine whether a stochastic or deterministic group testing method is most efficient. Chapter 4 proposes a novel method for modeling information flow patterns in conversations. The new method is demonstrated on a corpus of conversations between patients with advanced cancer and palliative care clinicians. The results confirm expected patterns in conversational discourse (validating the model) and provide novel insights about the normative patterns of information flow in serious illness conversation, and how these patterns are disrupted by the expression of distressing emotions. We conclude this dissertation in Chapter 5, with a summary of original contributions and ideas for future work, some with preliminary findings, on each of these problems.

# Bibliography

[1] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Petropolitanae*, pages 128–140, 1741.

[2] National Research Council et al. *Network Science*. National Academies Press,

2006.

[3] Ulrik Brandes, Garry Robins, Ann McCranie, and Stanley Wasserman. What is network science? *Network Science*, 1(1):1–15, 2013.

[4] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[5] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[6] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[7] Ann E Krause, Kenneth A Frank, Doran M Mason, Robert E Ulanowicz, and William W Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, 2003.

[8] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

[9] Jordi Bascompte, Pedro Jordano, Carlos J Melián, and Jens M Olesen. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387, 2003.

[10] Jens M Olesen, Jordi Bascompte, Yoko L Dupont, and Pedro Jordano. The

modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891–19896, 2007.

[11] Stanley Wasserman and Katherine Faust. Social network analysis in the social and behavioral sciences. *Social Network Analysis: Methods and Applications*, 1994:1–27, 1994.

[12] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[13] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.

[14] Alfredo Cuzzocrea, Alexis Papadimitriou, Dimitrios Katsaros, and Yannis Manolopoulos. Edge betweenness centrality: A novel algorithm for qos-based topology control over wireless sensor networks. *Journal of Network and Computer Applications*, 35(4):1210–1217, 2012.

[15] Kenneth L Calvert, Matthew B Doar, and Ellen W Zegura. Modeling internet topology. *IEEE Communications Magazine*, 35(6):160–163, 1997.

[16] Alexei Vázquez, Romualdo Pastor-Satorras, and Alessandro Vespignani. Large-scale topological and dynamical properties of the internet. *Physical Review E*, 65(6):066130, 2002.

[17] Michael GH Bell and Yasunori Iida. *Transportation Network Analysis*. 1997.

[18] Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network:

a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.

[19] Charles J Colbourn. Reliability issues in telecommunications network planning. In *Telecommunications Network Planning*, pages 135–146. Springer, 1999.

[20] Monojit Choudhury and Animesh Mukherjee. The structure and dynamics of linguistic networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer, 2009.

[21] Jin Cong and Haitao Liu. Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598–618, 2014.

[22] George Constable and Bob Somerville. *A Century of Innovation: Twenty Engineering Achievements that Transformed our Lives*. Joseph Henry Press, 2003.

[23] Qiming Chen, Chuanwen Jiang, Wenzheng Qiu, and James D McCalley. Probability models for estimating the probabilities of cascading outages in high-voltage transmission network. *IEEE Transactions on Power Systems*, 21(3):1423, 2006.

[24] 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.

[25] D.E. Newman, B.A. Carreras, V.E. Lynch, and I. Dobson. Exploring complex systems aspects of blackout risk and mitigation. *IEEE Transactions on Reliability*, 60(1):134 –143, 2011.

[26] EG Tiedemann. Channel access protocols for half duplex satellite terminals. In *IEEE Military Communications Conference,'Bridging the Gap. Interoperability, Survivability, Security'*, pages 463–469. IEEE, 1989.

[27] Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial Group Testing and its Applications*, volume 12. World Scientific, 2000.

[28] Dingzhu Du and Frank Hwang. *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

[29] Michael Kai Hourfar, Anna Themann, Markus Eickmann, Pilaipan Puthavathana, Thomas Laue, Erhard Seifried, and Michael Schmidt. Blood screening for influenza. *Emerging Infectious Diseases*, 13(7):1081, 2007.

[30] John R Papp, Julius Schachter, Charlotte A Gaydos, and Barbara Van Der Pol. Recommendations for the laboratory-based detection of chlamydia trachomatis and neisseria gonorrhoeae - 2014. *MMWR. Recommendations and Reports: Morbidity and mortality weekly report. Recommendations and Reports/Centers for Disease Control*, 63:1, 2014.

[31] Sun Bean Kim, Hye Won Kim, Hyon-Suk Kim, Hea Won Ann, Jae Kyoung Kim, Heun Choi, Min Hyung Kim, Je Eun Song, Jin Young Ahn, Nam Su Ku, et al. Pooled nucleic acid testing to identify antiretroviral treatment failure during hiv infection in seoul, south korea. *Scandinavian Journal of Infectious Diseases*, 46(2):136–140, 2014.

[32] Smriti Mallapaty. The mathematical strategy that could transform coronavirus testing. *Nature*, 583(7817):504–505, 2020.

[33] Christian Gollier and Olivier Gossner. Group testing against covid-19. *Covid Economics*, 2, 2020.

[34] Pooya Rezaei, Paul DH Hines, and Margaret Eppstein. Estimating cascading failure risk: Comparing monte carlo sampling and random chemistry. In *PES General Meeting | Conference & Exposition, 2014 IEEE*, pages 1–5. IEEE, 2014.

[35] Pooya Rezaei, Paul DH Hines, and Margaret J Eppstein. Estimating cascading failure risk with random chemistry. *IEEE Transactions on Power Systems*, 30(5):2726–2735, 2015.

[36] Steven C Amstrup, Trent L McDonald, and Bryan FJ Manly. *Handbook of Capture-Recapture Analysis*. Princeton University Press, 2010.

[37] RB O'hara. Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, pages 375–386, 2005.

[38] Chun-Huo Chiu, Yi-Ting Wang, Bruno A Walther, and Anne Chao. An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, 70(3):671–682, 2014.

[39] Anne Chao and Chun-Huo Chiu. Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, pages 1–26, 2014.

[40] Kenneth P Burnham and W Scott Overton. Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5):927–936, 1979.

[41] James F Heltshe and Nancy E Forrester. Estimating species richness using the jackknife procedure. *Biometrics*, pages 1–11, 1983.

[42] Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417):210–217, 1992.

[43] Anne Chao and Lou Jost. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547, 2012.

[44] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.

[45] Patrick Heneka and Bodo Ruck. A damage model for the assessment of storm damage to buildings. *Engineering Structures*, 30(12):3603–3609, 2008.

[46] 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.

[47] John D Musa and Kazuhira Okumoto. A logarithmic poisson execution time model for software reliability measurement. In *Proceedings of the 7th International Conference on Software Engineering*, pages 230–238. Citeseer, 1984.

[48] Valdo Durrleman, Ashkan Nikeghbali, and Thierry Roncalli. Which copula is the right one? *Available at SSRN 1032545*, 2000.

[49] Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle, and Tomasz Rychlik. *Copula Theory and its Applications*, volume 198. Springer, 2010.

[50] Fabrizio Durante and Carlo Sempi. *Principles of Copula Theory*. CRC press, 2015.

[51] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

[52] Lan Zhang and Vijay P Singh. *Copulas and Their Applications in Water Resources Engineering*. Cambridge University Press, 2019.

[53] Jérôme Lapuyade-Lahorgue, Jing-Hao Xue, and Su Ruan. Segmenting multi-source images using hidden markov fields with copula-based multivariate statistical distributions. *IEEE Transactions on Image Processing*, 26(7):3187–3195, 2017.

[54] Rong-Gang Cong and Mark Brady. The interdependence between rainfall and temperature: copula analyses. *The Scientific World Journal*, 2012, 2012.

[55] Patrick Laux, S Vogl, W Qiu, Hans Richard Knoche, and Harald Kunstmann. Copula-based statistical refinement of precipitation in rcm simulations over complex terrain. *Hydrology and Earth System Sciences*, 15:2401–2419.

[56] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier, 1978.

[57] Andreas Wulvik, Matilde Bisballe Jensen, and Martin Steinert. Temporal static visualisation of transcripts for pre-analysis of video material: Identifying modes of information sharing. In *Analysing Design Thinking: Studies of Cross-Cultural Co-Creation*. CRC Press, 2017.

[58] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[59] Howell Tong. Determination of the order of a markov chain by akaike's information criterion. *Journal of Applied Probability*, 12(3):488–497, 1975.

[60] Richard W Katz. On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249, 1981.

[61] Jennifer Pohle, Roland Langrock, Floris M van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):270–293, 2017.

[62] Ron Begleiter, Ran El-Yaniv, and Golan Yona. On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421, 2004.

[63] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-based Recognition*, pages 227–243. Springer, 1997.

[64] Gernot A Fink. *Markov Models for Pattern Recognition: From Theory to Applications*. Springer Science & Business Media, 2014.

[65] Wongyu Cho, Seong-Whan Lee, and Jin H Kim. Modeling and recognition of cursive words with hidden markov models. *Pattern Recognition*, 28(12):1941–1953, 1995.

[66] Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, and Chris Sander. Predicting protein structure using hidden markov models. *Proteins: Structure, Function, and Bioinformatics*, 29(S1):134–139, 1997.

[67] Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):121–125, 1999.

[68] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001.

[69] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 2, pages II–1. IEEE, 2003.

[70] Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[71] Scott M Thede and Mary Harper. A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 175–182, 1999.

# Chapter 2

# Risk of Cascading Blackouts Given Correlated Component Outages

## Abstract

Cascading blackouts typically occur when nearly simultaneous outages occur in $k$ out of $N$ components in a power system, triggering subsequent failures that propagate through the network and cause significant load shedding. While large cascades are rare, their impact can be catastrophic, so quantifying their risk is important for grid planning and operation. A common assumption in previous approaches to quantifying such risk is that the $k$ initiating component outages are statistically independent events. However, when triggered by a common exogenous cause, initiating outages may actually be correlated. Here, copula analysis is used to quantify the impact of correlation of initiating outages on the risk of cascading failure. The method is demonstrated on two test cases; a 2383-bus model of the Polish grid under varying load conditions and a synthetic 10,000-bus model based on the geography of the Western US. The large size of the Western US test case required development of new approaches for bounding an estimate of the total number of $N-3$ blackout-causing contingencies. The results suggest that both risk of cascading failure, and the relative contribution of higher order contingencies, increase as a function of spatial correlation in component failures.

## 2.1 Introduction

Cascading power failures are typically initiated when a small number of $k$ components in a power system of $N$ components disconnect nearly simultaneously, and the subsequent rerouting of power flow triggers additional component outages. This process continues until the system reaches a state of equilibrium. While most cascades do not propagate extensively throughout the network, the rare cases when they do can cause massive blackouts affecting millions of people. Due to their vast size and substantial social and economic costs, the risk they pose to power systems is significant [1, 2, 3].

Networks with heterogeneous load profiles, such as power systems, are particularly prone to cascades; without the right precautions, even a single node may trigger a cascade [4]. To mitigate the risk posed by cascading failure, power systems are required to operate such that no single component outage will cause a cascade (so-called $N$-1 security). While grid planners and operators are now also obligated to consider the risk of cascading failure due to multiple contingencies ($k > 1$) [5], it is not yet clear how to estimate this risk. For brevity, minimal $N - k$ contingencies that result in a cascading blackout are referred to as "malignancies", while contingencies that do not cause a blackout are referred to as "benign" [6]. By "minimal", we mean that no smaller subset of outages results in a blackout. Analysis of high order malignancies is challenging due to the nonlinear ways in which cascades propagate, the vast number of $N - k$ malignancies, and the combinatorial search space of possible contingencies.

In addition to helping to quantify risk of cascading failure, studying $N - k$ malignancies may potentially inform mitigating actions. For example, prior research into a simple model of cascading overloads in communication networks [7] suggests

that the intentional removal of key components directly after initiating sets of outages may reduce the size of subsequent cascades. In a power system model of the Polish grid, optimally dispatching generation assuming a 50% reduction in line limits on the 3 branches that contribute the most to the risk of cascades from $N-2$ malignancies dramatically reduces the overall risk of cascading failure with only a modest increase in operational costs [8].

Many previous approaches to cascading failure risk analysis (including our own) assumed initiating component outages to be independent events [1, 8, 9, 10, 11]. However, $N-k$ malignancies triggered by the same exogenous event, or "common cause", represent a significant source of risk to power systems [12], and can result in spatial correlation in initiating outages. For example, extreme weather events can result in spatially correlated damage [13], protection system failures can sometimes cause multiple outages within a small geographic region [14], and terrorist attacks may be spatially localized, such as in the 2013 sniper attack on the Metcalf substation near San Jose, CA, where the perpetrators shot 17 transformers at the same substation [15]. Non-spatial attributes, such as component age, may also induce correlations in component failures [16, 17].

There is a dramatically increasing computational burden to assessing risk for $N-k$ malignancies as $k$ increases, so it is important to understand the degree to which higher-order ($k > 2$) malignancies contribute to risk, and thus how important it is to consider them in risk estimation. Even though there are many more $N-3$ than $N-2$ malignancies for any given system, when there is no correlation in initiating outages the probability of $N-3$ malignancies occurring is so much lower than that of $N-2$ malignancies that the impact of $N-3$ malignancies on risk is negligible [9].

However, as correlation in component outages increases, the impact of higher order malignancies on risk will increase. To what degree should risk analysis take into account the conditional probability of component failure, given a common cause?

There has been some prior work on ways to incorporate correlation into risk analysis. In [18], correlation was incorporated by assuming 100% correlation of outages within a fixed radius. In [13], spatial correlation was achieved by determining outage rates of lines adjacent to initial failures probabilistically, according to a Poisson process. In [19], a random field with spatial autocorrelation was used in a cascade model to assess risk from common-cause events. Others [20] have simulated the impact of hidden relay failures on cascading failure risk by allowing proximate lines to trip probabilistically.

Another approach to incorporating correlation into risk estimation is *via* copula analysis. Popularized in the field of finance [21], copulas have been used in a wide variety of disciplines to model the co-dependence of multiple variables [22, 23]. Within the realm of power systems, copulas are a popular tool for uncertainty analysis. They have been used in the modelling of stochastic generation, such as wind [24, 25, 26]. The impacts of variable infeeds on security assessment have also been considered using copulas [27]. Li [28] suggests copulas as a useful way to incorporate correlation between random variables in power systems risk analysis.

A flexible and generalizable approach to risk estimation given correlated component outages was presented in [29] and used to estimate risk due to $N-2$ malignancies in a 2383-bus model of the Polish grid. This paper extends that work in several significant ways including: (i) incorporating the effects of $N-3$ malignancies, (ii) studying how the risk due to $N-3$, relative to $N-2$, malignancies changes as a function of correlation in outage probabilities, and (iii) applying the method to a much larger and

more geographically realistic 10,000-bus test case, which necessitated (iv) development of new methods for estimating the total number of $N-3$ malignancies.

This paper is organized as follows: methods for risk estimation using samples of $N-k$ malignancies, and the computationally efficient "Random Chemistry" (RC) sampling method used in this work, are reviewed in Sections 2.2.1 and 2.2.2, respectively. In Section 2.2.3 a method using copula analysis to incorporate initiating outage correlations into risk estimation is presented, and in Section 2.2.4 an approach to quantifying distance between transmission lines, when considering spatial correlation, is described. The two test cases used to demonstrate the method are described in Section 2.2.5; new methods for bounding the total number of $N-3$ malignancies in large systems are described in Section 2.2.6. Results and discussion are presented in Sections 2.3 and 2.4, respectively.

## 2.2 Methods

### 2.2.1 Estimating Risk of Cascading Failure

This study uses the method for estimating risk of cascading failure from sampled $N-k$ malignancies presented in [9, 11], briefly reviewed below.

The risk due to a set of branches (transmission lines or transformers) $\omega$ can be calculated as [30]:

$$R_\omega = p_\omega s_\omega \tag{2.1}$$

where $p_\omega$ is the joint probability of the branches in $\omega$ failing and $s_\omega$ is the size of the resultant blackout. Note that $p_\omega$ is itself a function of $p_i$, the independent outage

probability for each branch $i \in \omega$, as well as any effect of correlation among branch outage probabilities (as further defined in Section 2.2.3).

Blackout size $s_\omega$ is quantified as the total power (MW) unserved due to load shedding. In this work, a cascading blackout is considered to have occurred when 5% or more of the total load is shed in DCSIMSEP, a simulator of cascading outages in power systems [31]. The risk posed to the system by all $N-k$ malignancies comprising branches $\omega$, for a given $k$, is then:

$$R_k = \sum_{\omega \in \Omega_k} R_\omega \tag{2.2}$$

where $\Omega_k$ is the complete set of all $N-k$ malignancies for the specified $k$. For realistically-sized power systems it is not computationally tractable to find the entire set $\Omega_k$ for $k > 2$. However, if $\Omega_k^{sampled} \subset \Omega_k$ is a large and representative subset of size $|\Omega_k^{sampled}|$, comprising all unique $N-k$ malignancies found by many iterations of some sampling strategy, and if the size of the complete set of $N-k$ malignancies $|\Omega_k|$ can be estimated, then risk $\hat{R}_k$ associated with $N-k$ malignancies, for a given $k$, can be estimated as follows:

$$\hat{R}_k = \frac{|\Omega_k|}{|\Omega_k^{sampled}|} \sum_{\omega \in \Omega_k^{sampled}} R_\omega \tag{2.3}$$

Estimating $|\Omega_k|$ for $k > 2$ on large systems is itself a very challenging problem, as discussed in Section 2.2.6.

Considering only malignancies with $k \leq k_{max}$ and assuming that non-minimal supersets of malignancies do not substantially change the amount of load shed, as

justified in [9], the risk of cascading failure can be approximated as:

$$\hat{R} = \sum_{k \in \{2..k_{max}\}} \hat{R}_k \tag{2.4}$$

In this work, $k_{max} = 3$.

### 2.2.2 Random Chemistry Sampling

For each $k$, there are $\binom{N}{k}$ possible $N-k$ contingencies, only a small proportion of which are malignancies. While exhaustive search may be feasible (albeit time consuming) for $k = 2$, it is computationally intractable for $k > 2$ in large power systems. Thus, many iterations of the Random Chemistry (RC) sampling method were used to efficiently identify large sets of $N - k$ malignancies in each test case. RC is a stochastic set size reduction algorithm for identifying a small minimal set of initiating events that trigger some outcome of interest [32], and was first applied for identifying $N - k$ malignancies in power systems in [6]. For the reader's convenience, the RC algorithm is briefly reviewed below.

The RC algorithm uses a subset reduction scheme $\{a_1, a_2, \ldots, a_{final}\}$. Subsets of size $a_1$ are randomly sampled from a universal set of $N$ system components until one such set is found that causes a blackout; if $a_1$ is relatively large, this typically requires few tries. A set of size $a_{i+1}$ is then randomly sub-sampled from the preceding set of size $a_i$ until a set is found that causes a blackout (or some maximum number of sub-samples is tried, in which case the algorithm aborts), and so on for each subsequent set size in the scheme.

If $a_{i+1} = a_i/c$, for some constant $c$, then the algorithm requires only $O(N \log N)$

time to identify a subset of size $a_{final}$. As in [6], we use $c = 2$ from $a_1$ down to subsets of size 20 and then use $c = 1.5$ down to $a_{final}$. A bottom-up brute force search of all subsets of a given size $k$ is subsequently applied (conducted in randomized order, starting from $k = 2$), exiting when the first minimal malignancy of size $k = \{2, 3, \ldots, a_{final}\}$ is identified.

Repeated sampling with independent RC trials is performed to compile large subsets of $N - k$ malignancies ($\Omega_k^{RC}$). Risk due to each $k \leq k_{max}$ is then calculated using $\Omega_k^{sampled} = \Omega_k^{RC}$ in (2.3) for estimating system risk with (2.4).

A comparison of risk estimation using RC sampling *vs.* Monte Carlo (MC) sampling on a model of the Polish power system at peak winter load [33] showed that the RC approach was at least two orders of magnitude faster than MC on this system, and did not introduce measurable bias into the estimate [9, 11]. However, these previous studies assumed branch outages were uncorrelated. Under that assumption, $N-3$ malignancies contribute relatively little to the risk, despite the fact that $|\Omega_3^{RC}| \gg |\Omega_2^{RC}|$, since their probability of occurrence is so much smaller than that of the $N - 2$ malignancies [9].

In this study, the universal set is assumed to comprise the set of $N$ branches in each test case, $a_{final} = 5$, and up to 20 sub-samples at each set size were allowed before aborting an RC trial, as in [9, 11]. The specific set reduction scheme used for each test case is given in Section 2.2.5. Note that $k_{max} < a_{final}$ because, with the number of RC trials performed, $|\Omega_k^{RC}|$ was insufficient for estimating $|\Omega_k|$ for $k > 3$.

### 2.2.3 Copula Analysis for Correlation

Copula functions couple multivariate distributions to the marginal distributions of individual variables [34]. Given a set of $k$ random variables, $\mathbf{X} = [X_1, X_2, \ldots, X_k]$,

where $\Pr(X_i \leq t_i)$ is the marginal probability that branch $i$ fails, for some threshold $t_i$, then

$$F_X(\mathbf{T}) = \Pr\left(\bigcap_{i=1}^{k} X_i \leq t_i\right) \tag{2.5}$$

where $\mathbf{T} = [t_1, t_2, \cdots, t_k]$, represents the joint probability that $k$ branches fail together. Without loss of generality, it is assumed that $t_i = 0$ for all $i$.

There are numerous classes of copula functions in popular use. For this demonstration of the method, a Gaussian copula was assumed, but alternative distributions may be assumed where appropriate. Here, it is assumed that the inverse stress on a transmission line $i$ is a univariate Gaussian random variable $X_i = \mathcal{N}(\mu_i, \sigma_i)$, with mean $\mu_i$ and standard deviation $\sigma_i$, with the cumulative distribution function:

$$F_{X_i}(x_i) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x_i - \mu_i}{\sigma_i\sqrt{2}}\right)\right] \tag{2.6}$$

Given the independent probability $p_i$ of branch $i$ going out, $\mu_i$ and $\sigma_i$ are chosen such that when $X_i < 0$, the branch goes out. In other words, $\mu_i$ and $\sigma_i$ are chosen such that $F_{X_i}(0) = p_i$ for each branch $i$. Without loss of generality, it is assumed that $\mu_i = 1, \forall i$, and each $\sigma_i$ is then solved for as follows:

$$\sigma_i = \frac{-1}{\operatorname{erf}^{-1}(2p_i - 1)\sqrt{2}} \tag{2.7}$$

A multivariate normal distribution $\mathbf{X} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_k]$ and covariance matrix $\mathbf{C}$ is then used as the copula function to couple these univariate marginal distributions (Fig. 2.1).

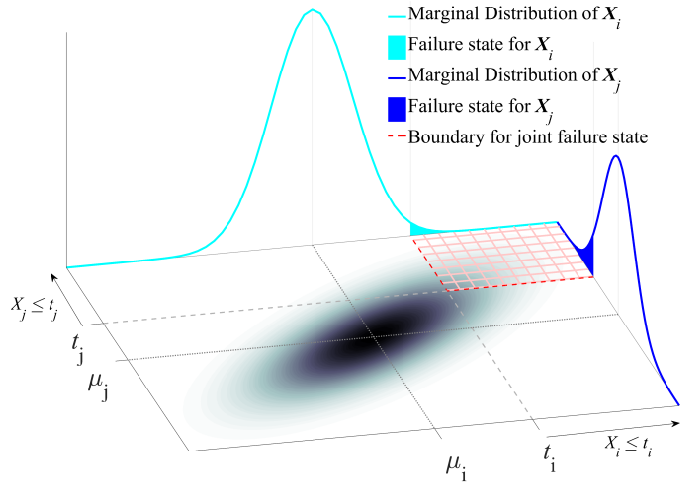



*Figure 2.1: A visual depiction of the copula method for two components $i$ and $j$ with hypothetical Gaussian distributions of some performance attributes, $X_i$ and $X_j$, which impact whether each component is operational or in a failure state. The curves on the vertical planes represent the marginal distributions of each component's attribute, with the shaded regions of these curves, $(X_i \leq t_i)$ and $(X_j \leq t_j)$, representing the failure state for each component. The shaded gradient on the horizontal plane represents the density of the joint distribution (copula) of the two variables, with darker shading representing higher probability density. The probability mass within the red hatched area represents the region of joint failure ($\mathbf{X} \leq [t_i, t_j]$), with the red dotted line depicting the boundaries of this region.*

In this study, it is assumed that the correlation between outages in branches $i$ and $j$ decays exponentially with the distance between them $d_{ij}$, according to:

$$\rho_{ij} = \rho_o e^{-d_{ij}/L} \tag{2.8}$$

where $\rho_o$ represents the maximum possible correlation coefficient (at distance zero) and $L$ represents the characteristic length, which controls the decay rate of the correlation; $L$ can be interpreted as the distance at which $\rho_{ij}$ reaches $\rho_o/e$ (i.e., $\approx 36.8\%$ of $\rho_o$) (Fig. 2.2).

Eq. (8) can be adjusted to represent a wide range of exogenous common cause events individually, or in combination, by adjusting the parameters $\rho_o$ and $L$ to

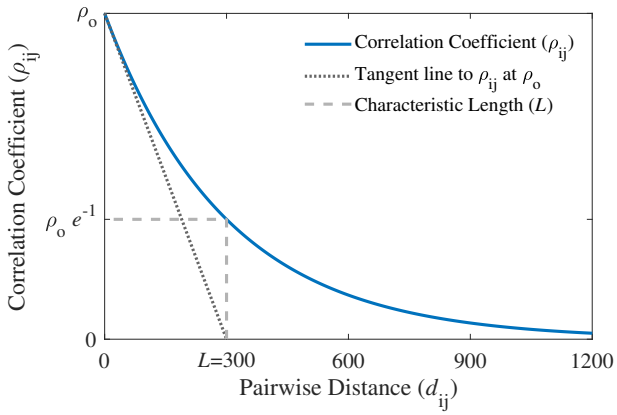



*Figure 2.2: Change in the correlation between two branches as a function of the distance between them, assuming* (2.8) *with characteristic length* $L = 300$ *km and correlation* $\rho_o$ *for branches that are 0 km apart.*

align with data for a particular set of threats. The exponential decay form captures the spatially decaying nature of earthquakes [35], and can approximately capture the impact of other threats that are likely to be geographically correlated, such as tornados [36] and hurricanes [37].

The resulting correlation coefficient $\rho_{ij}$ calculated by (2.8), and the standard deviations $\sigma_i$ and $\sigma_j$ calculated by (2.7), are used to calculate the pairwise covariance between branches $i$ and $j$ as:

$$\text{cov}(i, j) = \rho_{ij}\sigma_i\sigma_j \tag{2.9}$$

Using (2.9) to find each element of the covariance matrix $\mathbf{C}$, the probability density function of the multivariate normal distribution (2.10) is used to form the copula.

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{C}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2.10}$$

Integrating (2.10) over the region in the joint distribution that represents outages of all system components gives:

$$F_X(\mathbf{0}) = \int_{-\infty}^{0} \int_{-\infty}^{0} \cdots \int_{-\infty}^{0} f(x_1, x_2, \ldots, x_k)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots\, \mathrm{d}x_k \tag{2.11}$$

where $F_X(\mathbf{0})$ represents the joint outage probability of $k$ components $\Pr(\mathbf{X} \leq \mathbf{0})$. The multiple-integral in (2.11) represents the generalized solution for arbitrary $k$ and is equivalent to the cumulative distribution function of the multivariate normal distribution, which can be solved efficiently in MATLAB using methods described in [38]. In this work, $k \in \{2, 3\}$.

### 2.2.4 Defining Inter-Branch "Distance"

The definition of "distance" will vary based on the type of common cause threatening the system. Assuming spatial correlation in branch outages here, without loss of generality a modified version of the inter-branch distance metric defined in [29] is used. Branches are assumed to be straight lines between the buses that form their endpoints. Given branch $U$ with endpoints $(u_1, u_2)$ and branch $V$ with endpoints $(v_1, v_2)$, let the distance from $U$ to $V$ be defined as

$$\mathrm{Dist}(U, V) = \frac{\sum_{i=1}^{2} d(u_i, V) + \sum_{i=1}^{2} d(v_i, U)}{4} \tag{2.12}$$

where $d(v_i, U)$ is the minimum Euclidean distance from the point $v_i$ to the line segment $U = (u_1, u_2)$, calculated as:

$$d(v_i, U) = \begin{cases} ||v_i - u_1|| & t \leq 0 \\ ||v_i - (u_1 + tm)|| & 0 \leq t \leq 1 \\ ||v_i - u_2|| & t \geq 1 \end{cases} \tag{2.13}$$

where $m = u_2 - u_1$ and $t = \frac{(v_i - u_1) \cdot m}{||m||^2}$, as illustrated in Figure 2.3.

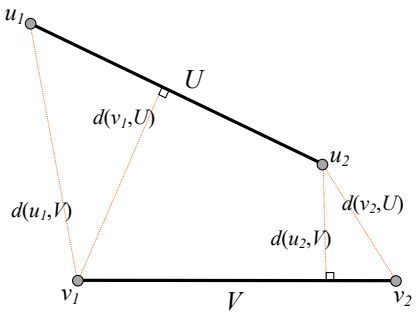



*Figure 2.3: Visual example for calculating the distance between branches $U$ and $V$ with endpoints $(u_1, u_2)$ and $(v_1, v_2)$, respectively.*

This formulation defines a semi-metric since the triangle inequality does not hold in some cases. However, all other formal requirements of a metric are met. Specifically:

1. $Dist(U, V) \geq 0$
2. $Dist(U, V) = Dist(V, U)$
3. $Dist(U, V) = 0 \iff U = V$

The third identity implies branches $U$ and $V$ share the same endpoints, thus are parallel. This $Dist(U, V)$ measure is consistent with what would be intuitively expected when

considering spatially correlated damage. For example, in Fig. 2.4, $Dist(A, B) > Dist(C, D)$ and $Dist(E, F) > Dist(G, H)$.

Using the $Dist(U, V)$ measure, it is apparent that branch pairs that form malignancies are much closer together than those of benign contingency pairs in both test cases (Fig. 2.5). This property will exacerbate the effects of spatial correlation on risk of cascading failure.

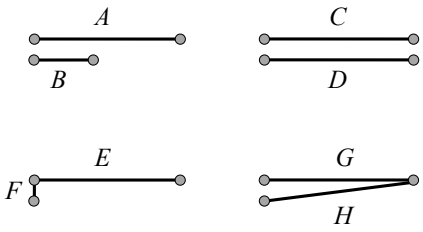



*Figure 2.4: Branch pairs used for pairwise distance examples described in the text.*

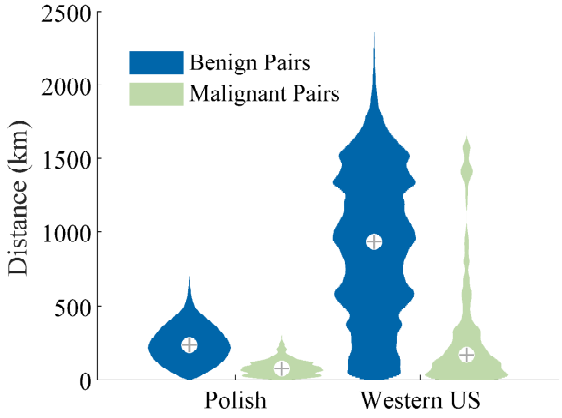



*Figure 2.5: Distance between the 540 and 564 branch pairs that form $N - 2$ malignancies in the Polish and Western US test cases* vs. *a random sample of 1,000,000 benign branch pairs from each test case. For clarity, medians are marked with crosshairs and each distribution has been independently normalized to the same maximum width.*

### 2.2.5 Case Studies

This risk estimation approach is demonstrated on two publicly available test cases, modeling the Polish and Western United States (US) transmission systems.

The Polish test case, examined in previous work on risk estimation [8, 9, 29], contains 2383 buses and 2896 branches at a peak winter load and is distributed with the MATPOWER simulation package [33]. The true spatial locations of branches and buses are not publicly available for this test case, so hypothetical locations were inferred based on a graph layout of the grid topology, assuming branches are straight lines between buses (Fig. 2.6). This layout was then scaled to $670 \times 670$ km, the approximate width/height of Poland, to simulate geographic distances. Some of the transmission lines were overloaded in the Polish test case provided by [33], so the adjusted base case described in [8] was used. Unless otherwise stated, references to the "Polish test case" refer to this adjusted base case. As in [9], different load levels were modeled in the Polish test case from 80% to 115% of the adjusted base case by multiplying all line loads by a scalar factor and then re-running the security constrained optimal power flow, to ensure the pre-contingency system at each load level is $N - 1$ secure.

The Western US test case is a synthetic network based on the footprint of the western Unites States and comes *via* the Electric Grid Test Case Repository [39]. This test case is much larger than the Polish test case, with 10,000 buses and 12,706 branches, and has a more realistic geographic layout (Fig. 2.7). As with the Polish test case, some transmission lines were overloaded for the Western US test case provided by [39], and so adjustments were made as described in [8]. Since the case did not

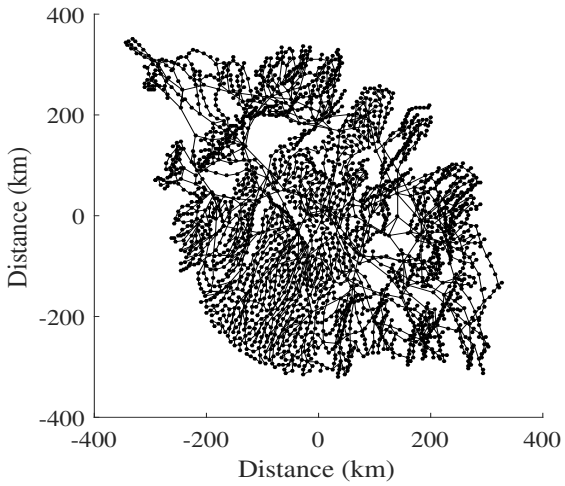



*Figure 2.6: Synthetic geographic layout of the Polish test case. Positionally, this layout is arbitrary and has been centered at (0,0), however, units were scaled so that the diameter of the geographic layout is roughly equal to that of Poland (in km).*

include short and long-term emergency flow limits ("RateB" and "RateC"), they were synthesized to be 110% and 150% of normal ("RateA") limits, respectively.

Independent branch outage rates were not available for either the Polish or Western US test cases. For the results presented here, all independent outage rates were set equal to the mean outage rate of 0.9158 hours per year provided by the RTS-96 test case [40]. These independent outage rates were deliberately assumed identical for all branches in order to more clearly elucidate the impact of spatial correlations in outage rates, as assessed using (2.4), for all combinations of $L \in \{0, 100, 200, 300\}$ km and $\rho_o \in \{0.00, 0.05, 0.10, 0.15\}$.

For the results shown here, the RC algorithm used the subset reduction scheme $\{80, 40, 20, 14, 10, 7, 5\}$ for the Polish model, as in [6, 8, 9, 11, 29]. For the larger Western US model, the initial RC subset size $a_1$ was raised to 320 to increase the probability that the initial subset causes a blackout; thus, the Western US test case used the subset reduction scheme $\{320, 160, 80, 40, 20, 14, 10, 7, 5\}$. We did 1,000,000

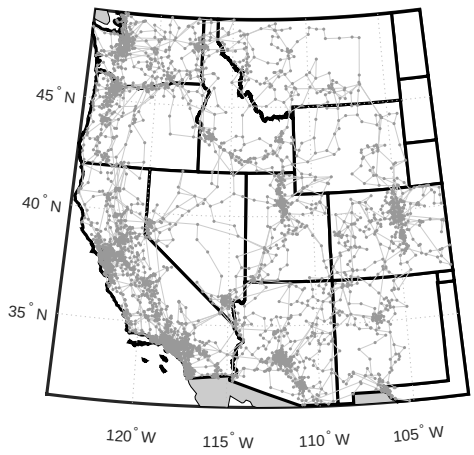



*Figure 2.7: Geographic layout of the synthetic 10,000 bus Western US test case.*

RC trials for the Polish test case and 704,400 RC trials for the Western US model; fewer RC trials were used for the larger test case because computation time was much higher than for the Polish model (averaging 9.5 seconds per RC trial for the Western US model *vs.* 2.35 seconds for the Polish model, on an Intel Core i5-3470 CPU @ 3.2GHz with 8 GB of RAM).

### 2.2.6 Estimating $|\Omega_k|$

As described in Section 2.2.1, this approach to risk estimation requires an estimate of the total number of $N-k$ malignancies $|\Omega_k|$, for $k \leq k_{max}$. There was no need to estimate $|\Omega_2|$, since RC sampling identified the complete set of $N-2$ malignancies $\Omega_2$ in both test cases, as evidenced by the flattening in the accumulation curves (Fig. 2.8 (top)), and later verified through brute force search for the Polish test case. The set of unique $N-2$ malignancies was complete after only 5,090 non-unique $N-2$ malignancies had been found by RC sampling in the Polish test case (of 4,191,960

possible $N-2$ contingencies) and after only 9,364 non-unique $N-2$ malignancies had been found by RC sampling in the Western US test case (of 80,714,865 possible $N-2$ contingencies).

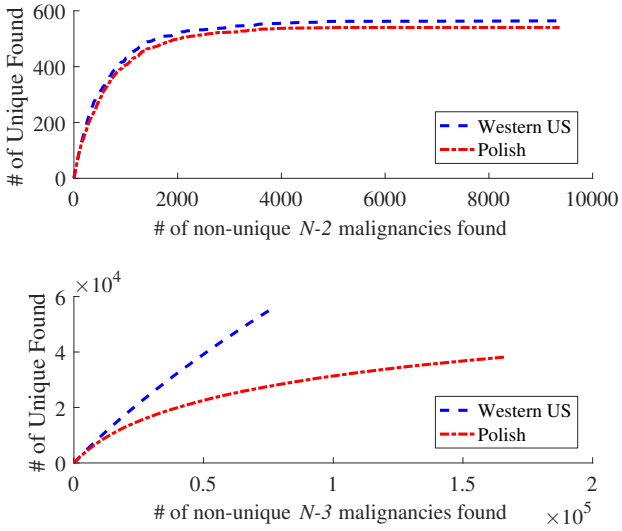



*Figure 2.8: (top) Accumulation curves for RC sampling of $N-2$ malignancies in the Polish and Western US test cases. In both cases, $|\Omega_2^{RC}| = |\Omega_2|$; (bottom) Accumulation curves for $N-3$ malignancies found by RC sampling in the Polish and Western US test cases. In both cases, $|\Omega_3^{RC}| \ll |\Omega_3|$.*

However, obtaining the entire set of N-3 malignancies is not computationally tractable in either test case, due to the sheer size of these sets. It was initially argued (incorrectly) in [6] that, if one has already identified $i$ of the $N-k$ malignancies using independent RC trials, then the probability that the next identified $N-k$ malignancy has not previously been found is $(|\Omega_k| - i)/|\Omega_k|$, so one could infer $|\Omega_k|$ from the observed frequency with which unique malignancies were found (assuming sufficient curvature in the accumulation curve). However, the assumption that independent RC trials uniformly sample from the $\Omega_k$ sets has since proven false. In subsequent studies it was discovered that the accumulation curves were not exponential (as they would be if the sampling were uniform), but could be better fit with a 4-parameter

exponential Weibull curve to estimate $|\Omega_k|$ [9]. While this non-linear curve-fitting approach works for estimating $|\Omega_3|$ in the Polish test case, the Western US test case is so much larger that there is insufficient curvature in the $N-3$ accumulation curve (Fig. 2.8 (bottom)) to reliably fit a curve.

It has previously been noted that the frequency of occurrence of individual branches in $N-2$ malignancies is heavy-tailed [6, 41]. A similarly heavy-tailed distribution is apparent in the distribution of occurrences of specific branch pairs in $N-3$ malignancies, in both the Polish and Western US (Fig. 2.9).

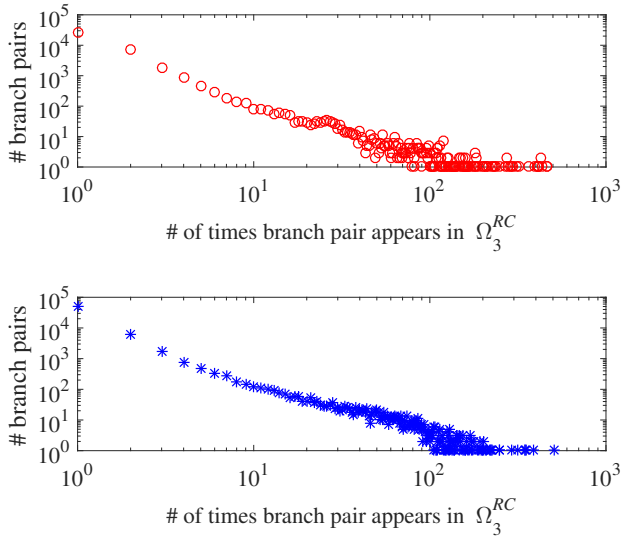



*Figure 2.9: Histograms of the number of occurrences of distinct branch pairs in unique $N-3$ malignancies found* via *RC in (top) the Polish test case and (bottom) the Western US test case.*

Further examination reveals that the set reduction scheme used in RC does, indeed, introduce a sampling bias when sampling from such heavy-tailed distributions. Specifically, the branch pairs that appear in disproportionately large numbers of $N-3$ malignancies are systematically under-sampled by the RC algorithm. To illustrate this, the 20 most frequently occurring branch pairs found in $|\Omega_3^{RC}|$ for the Western US

test case were selected, and a brute force search of all possible $N-3$ contingencies that included each of these top 20 branch pairs (requiring $O(N)$ computation time for each branch pair) was performed. In Fig. 2.10, the proportion of $N-3$ malignancies found by RC that contain each of these branch pairs is plotted as a function of the observed number of occurrences of the branch pairs in $\Omega_3^{RC}$. A clear negative trend is present, with $N-3$ malignancies containing the most frequently occurring branch pairs severely under-sampled relative to $N-3$ malignancies containing less frequently occurring branch pairs. While a thorough explanation of the causes of this sampling bias are beyond the scope of this paper, here the bias is exploited to estimate both lower and upper bounds on $|\Omega_3|$. (It is worth noting that, as $|\Omega_3^{RC}|$ approaches $|\Omega_3|$, the sampling bias of branch pairs found in $N-3$ malignancies decreases. However, for large networks such as the Western US test case, it is not computationally feasible to sample more than a small fraction of the $N-3$ malignancies, so the sampling bias remains high.)

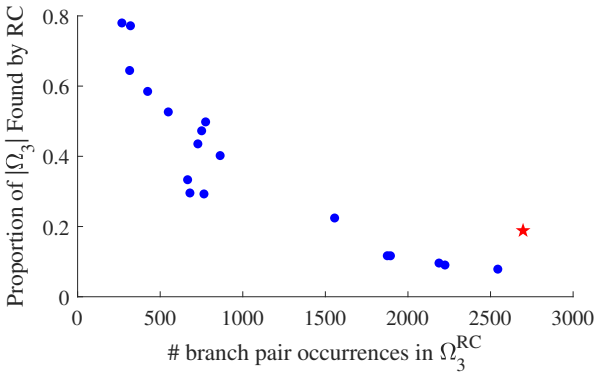



*Figure 2.10: The relation between number of occurrences of specific branch pairs in $N-3$ malignancies found by RC (x-axis) and the proportion RC has found of all $N-3$ malignancies that include those pairs (y-axis), for the Western US test case. Only the 20 most frequently occurring branch pairs are shown, with the star indicating the branch pair that occurred most frequently in $\Omega_3^{RC}$ ($Pair_{max}$).*

Given that sampling probabilities are unequal, this problem is analogous to the common conservation biology task of estimating population sizes *via* mark-and-recapture surveys in closed populations with heterogeneous sampling probabilities. There are numerous techniques that have been developed for this kind of problem [42]. Here, Chao's method [43] is used, because it is known to be particularly robust to heterogeneous sampling probabilities. In the power system context, the "population" under consideration is $\Omega_k$, the set of all $N-k$ malignancies. Chao's estimate is calculated as $|\Omega_3|^{Chao} = |\Omega_3^{RC}| + n_1^2/(2n_2)$, where $n_1$ is the number of $N-3$ malignancies found exactly once by RC sampling and $n_2$ is the number of $N-3$ malignancies found exactly twice by RC sampling. Chao's method produces a lower-bound on the population size within a fixed confidence interval [43], so it is assumed that $|\Omega_3|^{Chao} \leq |\Omega_3|$.

An upper-bound on $|\Omega_3|$ can be estimated by taking advantage of the two observations demonstrated above: (i) certain branch pairs appear disproportionately often in $N-3$ malignancies (Fig. 2.9), and (ii) the most frequently occurring branch pairs are under-sampled, relative to less frequent branch pairs (Fig. 2.10). Based on these observations, the Random Chemistry Proportional (RCP) method is proposed as a way to estimate an upper bound on $|\Omega_3|$, as follows: (i) apply RC sampling for a sufficient number of trials such that the identity of the most frequently occurring branch-pair ($Pair_{max}$) in the $N-3$ malignancies of the growing set $\Omega_3^{RC}$ becomes stable (for the Western US test case, $Pair_{max}$, indicated by the star in Fig. 2.10, became stable after about 7000 RC trials), (ii) perform a brute force search of all possible $N-3$ contingencies that include $Pair_{max}$ (this requires only $O(N)$ simulations) to determine the true number of $N-3$ malignancies that include $Pair_{max}$, (iii) compute what proportion of all $N-3$ malignancies that include $Pair_{max}$ were found by RC sampling,

and finally (iv) we estimate the total number of $N-3$ malignancies (referred to as $|\Omega_3|^{RCP}$) by assuming that all other less-frequently occurring branch-pairs have found this same proportion of the total number of $N-3$ malignancies in which they occur. Assuming that $Pair_{max}$ is under-sampled, this method provides an overestimate, and hence an upper-bound, on $|\Omega_3|$; i.e., it is expected that $|\Omega_3|^{RCP} > |\Omega_3|$.

As the number of $N-3$ malignancies found by RC sampling increases, $|\Omega_3|^{Chao}$ and $|\Omega_3|^{RCP}$ can be seen to be converging (Fig. 2.11), thus increasing the confidence in these as lower and upper bounds on the true value of $|\Omega_3|$. Risk estimates are calculated for the rightmost values of $|\Omega_3|^{Chao}$ and $|\Omega_3|^{RCP}$ shown in Fig. 2.11, to obtain approximate bounds on risk due to $N-3$ malignancies for the Western US test case.

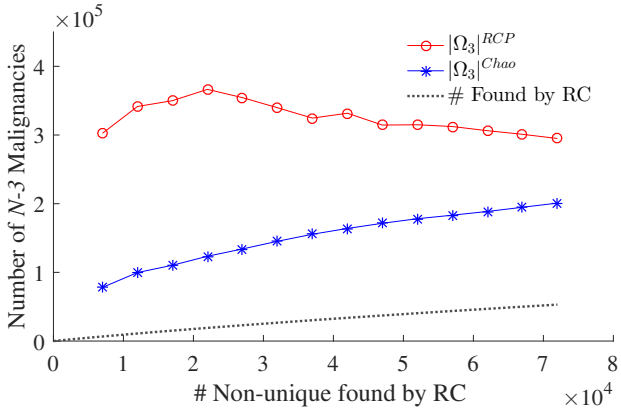



*Figure 2.11: The accumulation curve of $N-3$ malignancies found by RC is displayed below the lower-bound (Chao's method) and upper-bound (RCP method) estimates of $|\Omega_3|$ for the Western US test case.*

Similar approaches could conceivably be applied for estimating $|\Omega_k|$ for $k > 3$, however in this study $|\Omega_4^{RC}|$ and $|\Omega_5^{RC}|$ were insufficient to support this.

## 2.3 Results

### 2.3.1 Set sizes of $\Omega_2$ and $\Omega_3$

Brute force search was used to verify that RC sampling found all $N-2$ malignancies in the Polish base test case, with $|\Omega_2| = 540$. It is assumed that RC sampling also found all $N-2$ malignancies in the Western US test case with $|\Omega_2| = 564$, since the accumulation curve became flat (Fig. 2.8(top)). Using the non-linear curve-fitting method of [9], $|\Omega_3|$ is estimated to be $\approx 6.4 \times 10^4$ in the Polish test case at the base load. Using the Chao lower-bounding method [43] and the RCP upper-bounding method (described in Section 2.2.6), it is estimated that $2.0 \times 10^5 \leq |\Omega_3| < 2.9 \times 10^5$ in the Western US test case.

### 2.3.2 Impact of $N-2$ Correlation and Load Level on Risk

As shown in prior work [8, 9], the load levels on the Polish grid can greatly affect the vulnerability of the network to cascading power failure due to $N-2$ malignancies. As noted in [9], risk varies non-linearly and non-monotonically with load, in part due to variations in the proximity of generation to demand that result from optimal power flow dispatch at different load levels. Risk actually tends to drop at very high load levels because the security constrained optimal power flow results in more local generation, thus reducing the flow on critical long-distance transmission lines that can participate in many $N-k$ malignancies when overloaded. For a direct comparison to

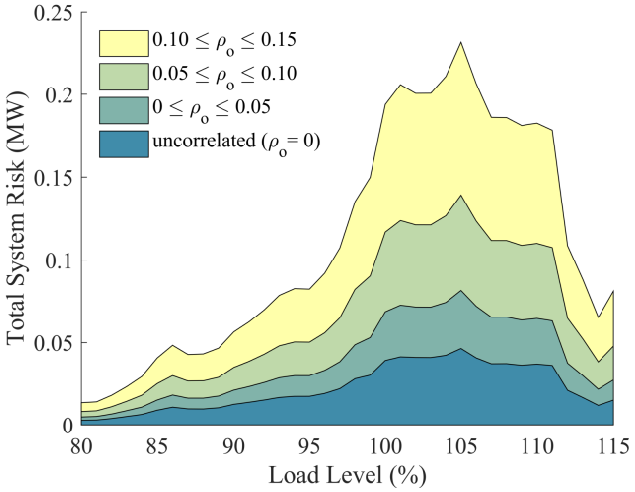



*Figure 2.12: Risk of cascading blackouts posed by spatially-correlated $N-2$ malignancies with a fixed characteristic length ($L = 300$ km) and varying values of maximum correlation $\rho_o$ for load levels from 80%-115% of the Polish base test case.*

the results presented in [9], the impact of spatial correlation in $N-2$ malignancies on risk was assessed as a function of load in the Polish test case.

Changes in the system risk as a function of load at $L = 300$ km (the longest characteristic correlation length tested) for 3 values of $\rho_o$ are illustrated in Fig. 2.12. Risk increased faster than linearly as a function of linearly increasing $\rho_o$, at each of the given load levels. The largest percentage increase in system risk occurred at load level 114%, while the greatest absolute increases in risk occur between load levels of 97%-112%, where there are the most $N-2$ malignancies. In general, while introducing correlation in initiating outages magnifies the risk of cascading blackouts, it does not fundamentally alter the overall shape of the risk curve as a function of load at $L = 300$ km.

When $\rho_o = 0.15$ (the largest $\rho_o$ tested) and $L$ was varied from 0 to 300 km, results were superficially similar to those in Fig. 2.12, in that higher correlation increases risk without changing the overall shape of the risk curve as a function of load. As was

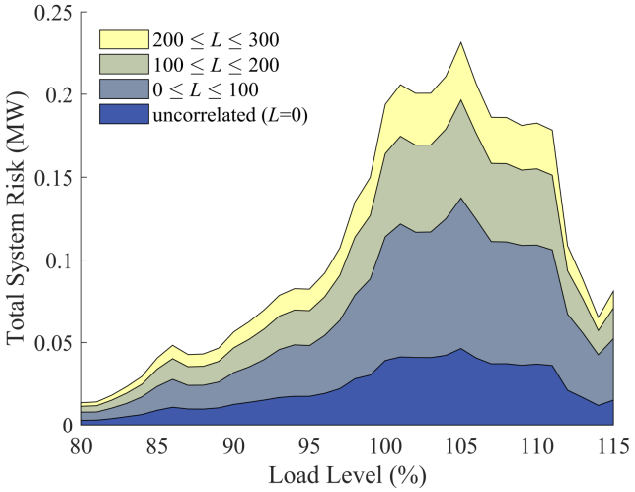



*Figure 2.13: Risk of cascading blackouts posed by spatially-correlated $N-2$ malignancies with a fixed maximum correlation ($\rho_o = 0.15$) and varying values of characteristic length $L$ (in km) for load levels from 80%-115% of the Polish base test case.*

the case when $L$ was fixed, the largest percentage increase in system risk was found to occur at load level 114%, and greatest absolute increases in risk occured between load levels of 97%-112%. However, in this case risk increases slower than linearly with linear increases in $L$, with the largest increases occurring for intermediate values of $L$ (Fig. 2.13). This occurs because increasing $L$ beyond a certain point has diminishing impact on correlation, as $L$ approaches the radius of the network.

The super-linear increases in risk as a function of $\rho_o$ and sub-linear increases in risk as a function of $L$ at the base load are clearly illustrated in Fig. 2.14.

### 2.3.3 Risk from $N-2$ and $N-3$ Malignancies

Risk of cascading blackouts posed by $N-2$ and $N-3$ malignancies was computed for both the Polish base test case and the Western US test case, over all values of $L$ and $\rho_o$ tested, using the set size estimates given in Sec. 2.3.1.

For the Polish test case (Table 2.1), the increase in estimated risk due to spatial

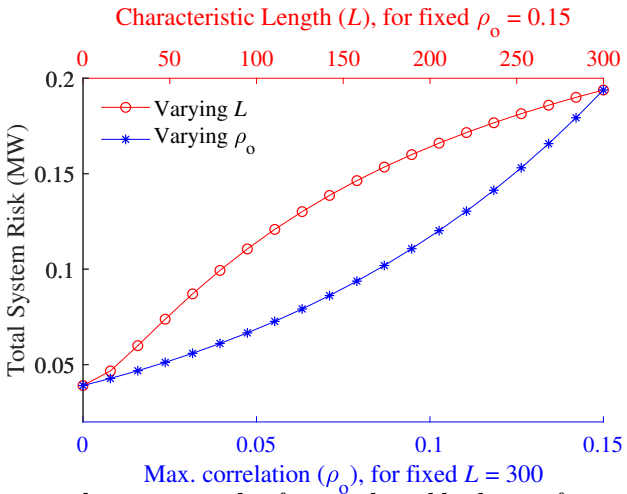



*Figure 2.14: Comparing change in risk of cascading blackouts for varying L (in km) with $\rho_o$ fixed at 0.15 (top x-axis)* vs. *varying $\rho_o$ with L fixed at 300 km (bottom x-axis) for the Polish base test case.*

correlation ranges from 149% for the most modest level of correlation tested ($L = 100$ km, $\rho_o = 0.05$) to 582% in the most extreme case tested ($L = 300$ km, $\rho_o = 0.15$), relative to the uncorrelated case.

*Table 2.1: Risk Attributable to $N-2$ and $N-3$ Malignancies in the Polish Test Case for Varying Levels of Spatial Correlation.*

| | $L$ (km) | | | |
|---|---|---|---|---|
| $\rho_o$ | 0 | 100 | 200 | 300 |
| 0.00 | 0.0394 | - | - | - |
| 0.05 | - | 0.0586 | 0.0675 | 0.0720 |
| 0.10 | - | 0.0876 | 0.1142 | 0.1290 |
| 0.15 | - | 0.1314 | 0.1918 | 0.2293 |

For the Western US test case, (Table 2.2), the increase in lower (upper) bounds on risk estimates varied from 129% (130%) for the most modest level of correlation tested ($L = 100$ km, $\rho_o = 0.05$) to 428% (456)% in the most extreme case tested ($L = 300$

*Table 2.2: Estimated Lower Bounds ($LB$) and Upper Bounds ($UB$) on Risk Attributable to $N-2$ and $N-3$ Malignancies in the Western US Test Case for Varying Levels of Correlation.*

| | | $L$ (km) | | | |
|---|---|---|---|---|---|
| $\rho_o$ | | 0 | 100 | 200 | 300 |
| 0.00 | $LB$ | 0.0654 | - | - | - |
| | $UB$ | 0.0665 | | | |
| 0.05 | $LB$ | - | 0.0846 | 0.0950 | 0.1019 |
| | $UB$ | | 0.0864 | 0.0974 | 0.1048 |
| 0.10 | $LB$ | - | 0.1148 | 0.1444 | 0.1654 |
| | $UB$ | | 0.1181 | 0.1502 | 0.1735 |
| 0.15 | $LB$ | - | 0.1631 | 0.2293 | 0.2801 |
| | $UB$ | | 0.1701 | 0.2445 | 0.3036 |

km, $\rho_o = 0.15$), relative to the uncorrelated case.

For both test cases, the general effect of $L$ and $\rho_o$ on risk that is described in Section 2.3.2 also holds in these results. That is, risk tends to grow faster than linearly with respect to $\rho_o$ and slower than linearly with respect to $L$. The larger proportionate increases in the Polish test case, relative to the Western US test case, occur because the average distance between branches in malignancies are shorter than in the Western US test case (Fig. 2.5), thus magnifying the impacts of spatial correlation.

### 2.3.4 Relative Risk of $N-2$ *vs.* $N-3$ Malignancies

It is expected that $N-3$ malignancies will contribute more to risk when there is spatial correlation in initiating outages, but it is not clear to what degree. There are several factors that could potentially disproportionately affect the impact of $N-3$ malignancies on risk when there is spatial correlation, relative to that of $N-2$ malignancies, including: (i) size of blackouts caused by $N-3$ *vs.* $N-2$ malignancies; (ii) the independent probability of branch outages in $N-3$ *vs.* $N-2$ malignancies; (iii) the distance between branches in $N-3$ *vs.* $N-2$ malignancies. These factors are each discussed in more detail below.

**Blackout Sizes**

If the sizes of blackouts resulting from $N-3$ malignancies were larger than $N-2$ malignancies, this could disproportionately increase the relative contribution of $N-3$ malignancies to risk when spatial correlation is present. However, we have observed that the sizes of cascading blackouts tend to follow similarly shaped distributions, independent of the number of component outages in the triggering event, due to similar patterns of network separation. This is illustrated by the distributions of blackout sizes (as estimated by DCSIMSEP) caused by all $N-k$ malignancies found by RC sampling, for $k \in \{2,3,4,5\}$, for both the Polish and Western US test cases (Fig. 2.15). In both test cases, the median blackout size for the identified $N-3$ malignancies was actually lower than those caused by the $N-2$ malignancies. Specifically, for the Polish test case, the median blackout size caused by $N-2$ malignancies was 7,624 MW *vs.*

3,372 MW for those caused by identified $N-3$ malignancies. In the Western US test case, the median blackout size from $N-2$ malignancies was 10,473 MW whereas from the $N-3$ malignancies it was 10,382 MW. These patterns and trends continue in the identified sets of $N-4$ and $N-5$ malignancies (Fig. 2.15).

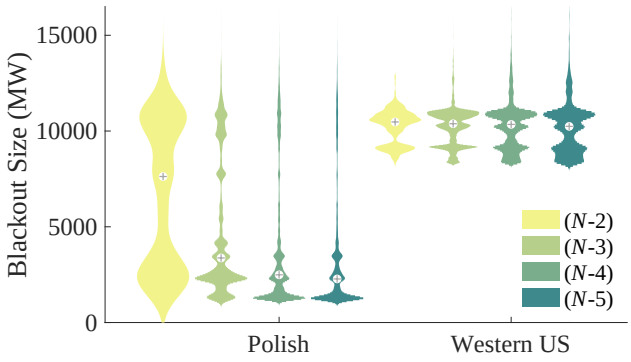



*Figure 2.15: Distributions of blackout sizes (in total MW load shed) caused by all $N-k$ malignancies ($2 \leq k \leq 5$) found by RC sampling, for the Polish and Western US test cases. For clarity, medians are marked with crosshairs and each distribution has been independently normalized to the same maximum width.*

### Independent Branch Outage Rates

In this study, independent outage rates were assumed to be homogeneous for all branches. However, in a real system the distribution of independent outage rates will be heterogeneous. If branches that are typically involved in $N-3$ malignancies are independently more likely to fail than those involved in $N-2$ malignancies, this could inflate the relative risk of $N-3$ malignancies when spatial correlation is present. While there is no obvious rationale for why this might be true, the observation that branches that occur frequently in $N-2$ malignancies also appear frequently in $N-3$ malignancies is indirect evidence against this. For example, in the Polish network, 8 of the 10 most frequently occurring branches in $N-2$ and $N-3$ malignancies

are shared, accounting for 44% and 24% of all $N-2$ and $N-3$ malignancies found, respectively. Likewise, for the Western US test case, 9 of the 10 most frequently occurring branches in $N-2$ malignancies are also in the top 10 most frequently occurring $N-3$ malignancies, accounting for 49% and 29% of all $N-2$ and $N-3$ malignancies, respectively.

### Distance Between Branches

The distance between branches in $N-3$ *vs.* $N-2$ malignancies will obviously impact the degree to which spatial correlation will increase their relative contributions to risk. In both the Polish and Western US test cases, median distances between all pairs of branches occurring in identified $N-k$ malignancies increases with $k \in \{2,3,4,5\}$ (Fig. 2.16). Specifically, in the Polish test case the median distances between pairs of branches were 76.1 km and 121.8 km, in $N-2$ and $N-3$ malignancies, respectively; in the Western US test case the medians were 169.4 km and 494.6 km in the $N-2$ and $N-3$ malignancies, respectively. This helps to mitigate the increase in relative risk from $N-3$ *vs.* $N-2$ malignancies that occurs as a result of spatial correlation.

### Comparing Relative Risk with Correlation

For the Polish test case, $< 1\%$ of risk can be attributed to $N-3$ malignancies when there is no correlation whereas under the highest level of correlation considered ($L = 300$ km, $\rho_o = 0.15$), the share of risk associated with $N-3$ malignancies rises to around 9% (Fig. 2.17). Similarly, for the Western US test case, $N-3$ malignancies account for 3%-5% of risk when there is no correlation, but between 16%-24% under the maximal correlation ($L = 300$ km, $\rho_o = 0.15$) considered (Fig. 2.18).

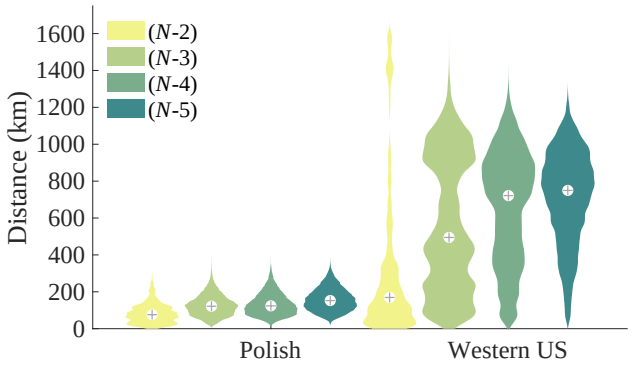



*Figure 2.16: Distributions of pairwise distances among branches in all $N-k$ malignancies ($2 \leq k \leq 5$) identified by RC sampling, for the Polish and Western US test cases. For clarity, medians are marked with crosshairs and each distribution has been independently normalized to the same maximum width.*

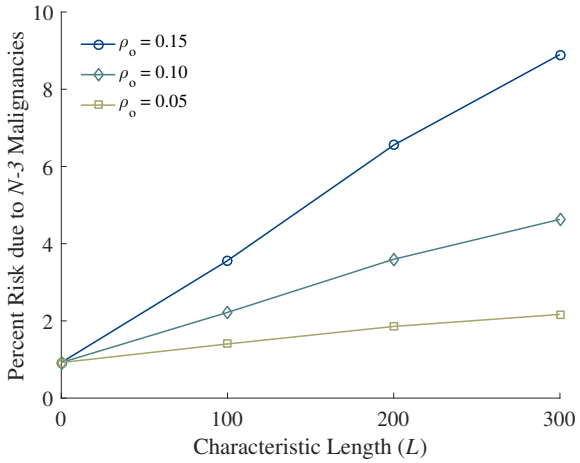



*Figure 2.17: Estimated percentage of risk attributable to $N-3$ malignancies* vs. *$N-2$ malignancies for the Polish Test Case under varying levels of correlation, including all combinations of $L \in \{0, 100, 200, 300\}$ km and $\rho_o \in \{0, 0.05, 0.10, 0.15\}$.*

## 2.4 Discussion

Previous research into cascading failure risk demonstrated that $N-3$ malignancies constitute a relatively low proportion of risk compared to $N-2$ malignancies, assuming initiating branch outage independence [9]. This suggests that, if initiating

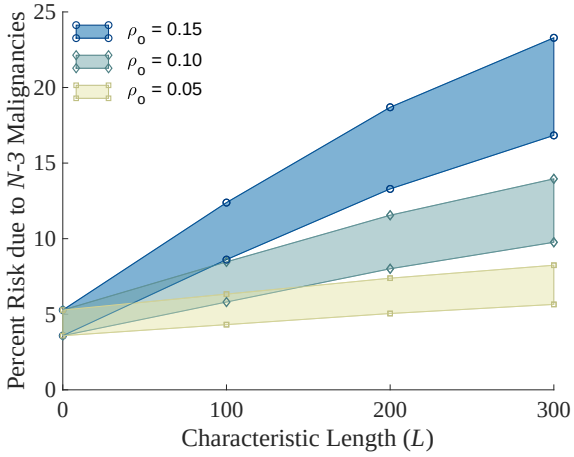



*Figure 2.18: Shaded regions represent bounded estimates on the percentage of risk attributable to $N-3$ malignancies* vs. *$N-2$ malignancies for the Western US Test Case under varying levels of correlation, including all combinations of $L \in \{0, 100, 200, 300\}$ km and $\rho_o \in \{0, 0.05, 0.10, 0.15\}$.*

outages are generally caused by independent events, limiting risk analysis to the more computationally tractable $N-2$ malignancies may be sufficient to capture the majority of risk. However, in reality, common causes such as relay failures, weather disturbances, earthquakes, fire, or spatially localized terrorist attacks may trigger multiple near-simultaneous outages in geographic proximity that could potentially result in cascading blackouts. For these cases an assumption of independence will under-estimate cascading blackout risk.

This paper presents a method that uses copula analysis as a flexible, customizable approach for incorporating correlation into risk calculations, building on preliminary work presented in [29]. The impact of spatial correlation in $N-2$ and $N-3$ initiating outages on risk of cascading blackouts is assessed in the Polish test case as well as the much larger, and geographically more realistic, Western US test case.

The Western US test case has over four times as many branches as the Polish

test case, and the increased computational cost of performing cascade simulations combined with the substantially larger search space of $N-3$ contingencies rendered previously developed methods [6, 9] for estimating the set size of $N-3$ malignancies ineffective. Thus, extending the approach to include risk from $N-3$ malignancies in the Western US test case required new methods to estimate lower- and upper-bounds on the total number of $N-k$ malignancies, for $k=3$.

The results indicate that when spatial correlation is present in initiating outages, the relative contribution of $N-3$ malignancies to risk of cascading blackouts increases, although the increase is partially mitigated by the fact that pairwise distances between branches in $N-3$ malignancies are greater than in $N-2$ malignancies.

It is expected that the impact of even higher-order malignancies will similarly increase with increasing spatial correlation, even though median pairwise distances between branches in malignancies continue to increase with $k$. In principle, the approaches to estimating lower and upper bounds on $|\Omega_3|$ presented here for the Western US test case could also be applied for estimating $|\Omega_k|$ for $k>3$, given a sufficient number of RC trials. We are currently exploring these and other methods on large synthetic networks to establish their limitations as a function of network size and functional heterogeneity. While ignoring higher-order $N-k$ malignancies when component outages are correlated will likely underestimate the magnitude of the risk of cascading failures, estimating $|\Omega_k|$ for $k$ higher than 3 (or possibly 4) on large networks may not be computationally tractable, due to the sheer number of these high-order malignancies. Preliminary work indicates that, for the Western US test case, $|\Omega_4|$ may be at least two orders of magnitude higher than $|\Omega_3|$.

The lack of accurate data regarding independent transmission line outage rates

and the impacts of common cause events on those rates are an important limitation to applying these methods in practice. Some such data are available to industry through systems such as the NERC TADS database, but these data are typically unavailable for research purposes. Increasingly, efforts are being made to better predict how common-cause events such as weather-related events will impact the grid [44]. Such knowledge will inform specific applications of the generalized framework introduced herein.

The methods presented in this paper should work with any simulator (AC, DC, or even something more sophisticated like a full dynamics cascading failure simulator). However, more complicated simulation models require much larger input datasets and tuning these models to get accurate results is a longer process (see [45, 46] for illustrations of the challenges associated with dynamic modeling of cascading failure). For example, in order to get an accurate model from an AC or dynamic power system simulator one would need to accurately model all of the dynamic reactive power elements, such as synchronous condensers and switched capacitor banks, in order to get accurate results. The impact of these controls is that a system with large amounts of reactive support will act more like a DC model with uniform voltage, than an AC model without reactive support. Since the focus of this paper is primarily on the computational method for risk analysis, rather than precise power systems details, we have used a simulator based on the DC power flow. Based on our experience with more complicated simulation models, we do not expect that a more complicated simulator would produce qualitative differences in the results, although quantitative differences would result since there are more mechanisms of cascading in an AC power flow model.

Future work will study the impact of parametric choices such as distance metrics, correlation functions, and the underlying branch outage probability distributions. The method will also be applied to study the risk of cascading failure in interdependent networks, which are ubiquitous in human-engineered infrastructures [47]. For example, coupling between communication and power networks can substantially impact their robustness to cascading failures [48, 49]; the method presented herein could help to quantify the impact of this coupling on risk in the presence of correlated component outages. As new methods for measuring the risk of cascading failure in systems with correlated initiating event probabilities emerge, there will be a need for comparisons to better understand the relative computational efficiency and accuracy of these approaches.

# Bibliography

[1] Qiming Chen, Chuanwen Jiang, Wenzheng Qiu, and James D McCalley. Probability models for estimating the probabilities of cascading outages in high-voltage transmission network. *IEEE Transactions on Power Systems*, 21(3):1423, 2006.

[2] 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.

[3] D.E. Newman, B.A. Carreras, V.E. Lynch, and I. Dobson. Exploring complex systems aspects of blackout risk and mitigation. *IEEE Transactions on Reliability*, 60(1):134 –143, 2011.

[4] Adilson E Motter and Ying-Cheng Lai. Cascade-based attacks on complex networks. *Physical Review E*, 66(6):065102, 2002.

[5] Standard NERC. Top-004–2: Transmission operations. *North American Electric Reliability Corporation*, 2007.

[6] Margaret J Eppstein and Paul DH Hines. A "random chemistry" algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[7] Adilson E Motter. Cascade control and defense in complex networks. *Physical Review Letters*, 93(9):098701, 2004.

[8] Pooya Rezaei, Margaret J Eppstein, and Paul DH Hines. Rapid assessment, visualization, and mitigation of cascading failure risk in power systems. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 2748–2758. IEEE, 2015.

[9] Pooya Rezaei, Paul DH Hines, and Margaret J Eppstein. Estimating cascading failure risk with random chemistry. *IEEE Transactions on Power Systems*, 30(5):2726–2735, 2015.

[10] Klaus Köck, Herwig Renner, and Josef Stadler. Probabilistic cascading event risk assessment. In *Power Systems Computation Conference (PSCC), 2014*, pages 1–7. IEEE, 2014.

[11] Pooya Rezaei, Paul DH Hines, and Margaret Eppstein. Estimating cascading failure risk: Comparing monte carlo sampling and random chemistry. In *PES General Meeting | Conference & Exposition, 2014 IEEE*, pages 1–5. IEEE, 2014.

[12] Milorad Papic, Sudhir Agarwal, Ron N Allan, Roy Billinton, Chris J Dent, Svetlana Ekisheva, Daniel Gent, Kai Jiang, Wenyuan Li, Joydeep Mitra, et al. Research on Common-Mode and Dependent (CMD) outage events in power systems: A review. *IEEE Transactions on Power Systems*, 32(2):1528–1536, 2017.

[13] Ian Dobson, NichelleLe K Carrington, Kai Zhou, Zhaoyu Wang, Benjamin A Carreras, and José M Reynolds-Barredo. Exploring cascading outages and weather via processing historic data. In *Hawaii International Conference on System Sciences (HICSS-51)*, Jan 2017.

[14] Kai Jiang and Chanan Singh. New models and concepts for power system reliability evaluation including protection system failures. *IEEE Transactions on Power Systems*, 26(4):1845–1855, 2011.

[15] Rebecca Smith. Us risks national blackout from small-scale attack. *Wall Street Journal*, 12, 2014.

[16] Wenyuan Li. Incorporating aging failures in power system reliability evaluation. *IEEE Transactions on Power Systems*, 17(3):918–923, 2002.

[17] Abdullahi M Salman. Age-dependent fragility and life-cycle cost analysis of timber and steel distribution poles subjected to hurricanes. Master's thesis, Michigan Technological University, 2014.

[18] Andrey Bernstein, Daniel Bienstock, David Hay, Meric Uzunoglu, and Gil Zussman. Power grid vulnerability to geographically correlated failures - analysis and control implications. In *INFOCOM, 2014 Proceedings IEEE*, pages 2634–2642. IEEE, 2014.

[19] Anke Scherb, Luca Garrè, and Daniel Straub. Reliability and component importance in networks subject to spatially distributed hazards followed by cascading failures. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 3(2):021007, 2017.

[20] Jie Chen, James S Thorp, and Ian Dobson. Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model. *International Journal of Electrical Power & Energy Systems*, 27(4):318–326, 2005.

[21] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

[22] Arno Onken, Steffen Grünewälder, Matthias HJ Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS Computational Biology*, 5(11):e1000577, 2009.

[23] Christian Schoelzel and Petra Friederichs. Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15(5):761–772, 2008.

[24] Simeon Hagspiel, Antonis Papaemannouil, Matthias Schmid, and Göran Andersson. Copula-based modeling of stochastic wind power in europe and implications for the swiss power grid. *Applied energy*, 96:33–44, 2012.

[25] George Papaefthymiou and Pierre Pinson. Modeling of spatial dependence in wind power forecast uncertainty. In *Probabilistic Methods Applied to Power Systems,*

*2008. PMAPS'08. Proceedings of the 10th International Conference on*, pages 1–9. IEEE, 2008.

[26] George Papaefthymiou and Dorota Kurowicka. Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE Transactions on Power Systems*, 24(1):40–49, 2009.

[27] Martijn de Jong, Georgios Papaefthymiou, and Peter Palensky. A framework for incorporation of infeed uncertainty in power system risk-based security assessment. *IEEE Transactions on Power Systems*, 33(1):613–621, 2018.

[28] Wenyuan Li. *Risk assessment of power systems: models, methods, and applications*. John Wiley & Sons, 2014.

[29] Laurence A Clarfeld, Margaret J Eppstein, Paul DH Hines, and Eric M Hernandez. Assessing risk from cascading blackouts given correlated component failures. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7. IEEE, 2018.

[30] Marianna Vaiman, Keith Bell, Yousu Chen, Badrul Chowdhury, Ian Dobson, Paul Hines, Milorad Papic, Stephen Miller, and Pei Zhang. Risk assessment of cascading outages: Methodologies and challenges. *IEEE Transactions on Power Systems*, 27(2):631, 2012.

[31] Paul Hines and Pooya Rezaei. *Smart Grid Handbook*, chapter Cascading Failures in Power Systems. John Wiley & Sons, 2016.

[32] Margaret J Eppstein, Joshua L Payne, Bill C White, and Jason H Moore. Genomic mining for complex disease traits with "random chemistry". *Genetic Programming and Evolvable Machines*, 8(4):395–411, 2007.

[33] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, Robert John Thomas, et al. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011.

[34] Roger B Nelsen. *An introduction to copulas*. Springer, New York, 2010.

[35] Nirmal Jayaram and Jack W Baker. Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics*, 38(15):1687–1708, 2009.

[36] Joshua Wurman and Curtis R Alexander. The 30 may 1998 spencer, south dakota, storm. part ii: Comparison of observed damage and radar-derived winds in the tornadoes. *Monthly Weather Review*, 133(1):97–119, 2005.

[37] HE Willoughby, RWR Darling, and ME Rahn. Parametric representation of the primary hurricane vortex. part ii: A new family of sectionally continuous profiles. *Monthly Weather Review*, 134(4):1102–1120, 2006.

[38] Alan Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14(3):251–260, 2004.

[39] Adam B Birchfield, Ti Xu, Kathleen M Gegner, Komal S Shetye, and Thomas J Overbye. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on Power Systems*, 32(4):3258–3265, 2017.

[40] RTS Task Force. The ieee reliability test system-1996. *IEEE Trans. Power Syst*, 14(3):1010–1020, 1999.

[41] Yang Yang, Takashi Nishikawa, and Adilson E Motter. Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365):eaan3184, 2017.

[42] Steven C Amstrup, Trent L McDonald, and Bryan FJ Manly. *Handbook of capture-recapture analysis*. Princeton University Press, 2010.

[43] Anne Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791, 1987.

[44] Yang Liu and Jin Zhong. Risk assessment of power systems under extreme weather conditions-a review. In *PowerTech, 2017 IEEE Manchester*, pages 1–6. IEEE, 2017.

[45] Jiajia Song, Eduardo Cotilla-Sanchez, Goodarz Ghanavati, and Paul DH Hines. Dynamic modeling of cascading failure in power systems. *IEEE Transactions on Power Systems*, 31(3):2085–2095, 2016.

[46] Dmitry N Kosterev, Carson W Taylor, and William A Mittelstadt. Model validation for the august 10, 1996 wscc system outage. *IEEE Transactions on Power Systems*, 14(3):967–979, 1999.

[47] Steven M Rinaldi, James P Peerenboom, and Terrence K Kelly. Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems Magazine*, 21(6):11–25, 2001.

[48] Mert Korkali, Jason G Veneman, Brian F Tivnan, James P Bagrow, and Paul DH Hines. Reducing cascading failure risk by increasing infrastructure network interdependence. *Scientific Reports*, 7:44499, 2017.

[49] Zhenhao Chen, Jiajing Wu, Yongxiang Xia, and Xi Zhang. Robustness of interdependent power grids and communication networks: A complex network perspective. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(1):115–119, 2018.

# Chapter 3

# Group Testing for Efficiently Sampling Hypergraphs When Tests Have Variable Costs

## Abstract

In the group-testing literature, efficient algorithms have been developed to minimize the number of tests required to identify all minimal "defective" sub-groups embedded within a larger group, using deterministic group splitting with a generalized binary search. In a separate literature, researchers have used a stochastic group splitting approach to efficiently sample from the intractable number of minimal defective sets of outages in electrical power systems that trigger large cascading failures, a problem in which positive tests can be much more computationally costly than negative tests. In this work, we generate test problems with variable numbers of defective sets and a tunable positive:negative test cost ratio to compare the efficiency of deterministic and stochastic adaptive group splitting algorithms for identifying defective edges in hypergraphs. For both algorithms, we show that the optimal initial group size is a function of both the prevalence of defective sets and the positive:negative test cost ratio. We find that deterministic splitting requires fewer total tests but stochastic

splitting requires fewer positive tests, such that the relative efficiency of these two approaches depends on the positive:negative test cost ratio. We discuss some real-world applications where each of these algorithms is expected to outperform the other.

## 3.1 Introduction

The field of group testing is thought to have originated from a single report by Dorfman in which he proposed a novel method for efficiently screening soldiers for syphilis during World War II [1]. Dorfman suggested mixing together blood samples from multiple individuals so it would require just a single chemical test to determine if the pooled blood sample contained syphilitic antigen. If the test came back negative, it would indicate that none of the soldiers were infected, whereas a positive test result would require subsequent tests to determine which soldiers were infected. Although this initial proposal was never implemented, group testing has since been applied to solve combinatorial search problems in a variety of disciplines [2]. Group testing strategies are currently being explored to minimize the number of tests needed estimate the prevalence of Covid-19 in large swaths of the population (*e.g.*, [3, 4]). Finding optimal group testing strategies that minimize the number of tests required to identify "defective" individuals (or items, or minimal sets of items) has been a central focus in the robust group-testing literature that has emerged. An implicit assumption of this work is that the cost of positive tests is the same as the cost of negative tests. In this paper, we examine so-called "adaptive" group testing algorithms for finding defective edges in hypergraphs, where the results of previous tests are used to inform which tests to perform next.

A hypergraph is a generalization of an ordinary graph, in which individual edges

(so-called "hyperedges") can connect an arbitrary number of nodes. Specifically, define a hypergraph $G(V, E)$ where each node in $V$ represents one of the $N$ individuals or items of interest and each hyperedge in $E$ connects a set of $k$ nodes, or $k$-set. We consider the most general case, where there exists an edge in $E$ for every possible subset of $V$. When using group-testing to identify defective edges in a hypergraph, the aim is to identify minimal defective $k$-sets in $E$, where "minimal" means that no smaller subset is defective.

Searching for defective edges in a graph $G$ with $|E|$ edges using group testing was first considered by Aigner, who conjectured that no more than $\lceil \log_2 |E| \rceil + c$ tests (for some constant $c$) were required to find a single defective edge in $G$ [5]. This was later proven by Damaschke [6] and generalized to hypergraphs by Triesch [7]. The case of finding all $d > 1$ defective edges in a graph, where $d$ is known, was first addressed in [8]. For the case when $d$ is unknown, adaptive methods for finding all defective edges were proposed for graphs in [9] and extended to hypergraphs in [10, 11]. All of the aforementioned algorithms use deterministic splitting approaches.

One important real-world problem that can be framed as searching for defective hyperedges in a hypergraph is the identification of minimal sets of $k$ outages that trigger cascading failures in power systems, as a means of estimating overall risk of cascading failure. This real-world problem has several characteristics that have not traditionally been considered in the standard group-testing literature: (i) The number of minimal defective edges $d$ is so large that it is computationally intractable to identify them all, so sampling is required; (ii) $G$ is non-uniform (*i.e.*, $k$ is variable, in the $k$-sets defined by the hyperedges in $E$); (iii) the size of the defective sets sought is lower-bounded by $k_{min} = 2$, since power systems are operated such that no

single outage ($k = 1$) will result in a cascading failure; (iv) the size of the defective sets sought is often upper bounded by $k_{max}$, due to computational limitations; (v) tests can produce false negatives, wherein a set tests as non-defective even though it contains a defective subset, (*e.g.*, this can occur when a power grid is fragmented into disconnected but independently functioning "islands" [12]); and, most importantly, (vi) positive and negative tests may have very different costs.

The reason that test costs are so variable in the power systems problem is that simulating each step of a cascade accrues additional computational costs, and positive tests have larger cascades than negative tests, by definition. We illustrate an example of this difference in the costs of positive and negative tests using DCSIMSEP [13] (Fig. 3.1), an open-source DC simulator that has been used in studying cascading power outages [14, 15, 16, 17, 18, 19]. We note that a variety of types of DC and AC power flow solvers and cascade models could theoretically be used to simulate cascading failures in power grids, each with their own sets of advantages and disadvantages [20], [21], [22], [23], [24],[25], [26]. If one were to use a more sophisticated AC simulator, the relative difference between the computational cost of positive and negative tests would be expected to be even greater, because each step of an AC simulator will be more computationally costly than in a DC simulator [21].

In the power systems literature, one method that has been proposed to tackle the problem of efficiently finding minimal defective $k$-sets is the Random Chemistry (RC) algorithm [14, 15, 16, 17, 18, 19]. RC is a stochastic adaptive group testing approach. To our knowledge, deterministic group-testing approaches have not previously been applied to the power systems problem or other problems where the cost of positive and negative tests is unequal.

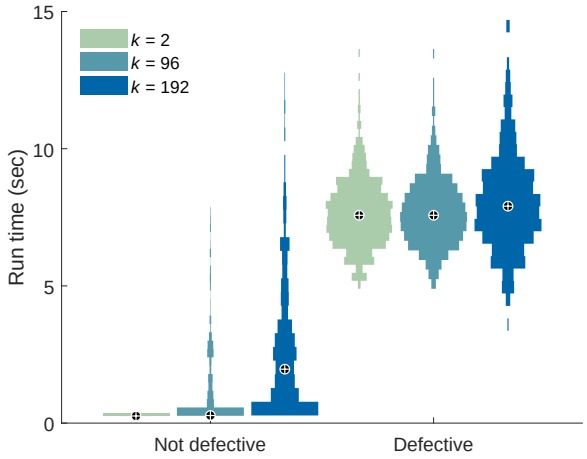



*Figure 3.1: Run times required using DCSIMSEP to test 500 random non-defective sets and 500 random defective sets, for each of set sizes $k \in \{2, 96, 192\}$, on a synthetic model of the Western U.S. power grid. To generate this data, a set was considered defective if there was at least 5% of the load shed in DCSIMSEP. For clarity, medians are marked with crosshairs and each distribution has been independently normalized to the same maximum width. See Sec. 3.2.3 for a description of the simulator and the test case.*

Although this work was originally motivated by the power systems problem described above, there are other potential applications that share many of these characteristics, including variable positive:negative test costs (see Sec. 3.4). Thus, the aim of this study is to explore the general question regarding the relative computational efficiency of deterministic and stochastic adaptive group-testing algorithms for sampling minimal defective sets, in problems where there are large and unknown minimal defective $k$-sets with variable $k$, the potential for false negatives exists, and in which the relative computational costs of positive and negative tests may differ.

To that end, we create test problems with different numbers of minimal defective hyperedges, low (but non-zero) frequencies of false negatives, and tunable positive:negative test costs. We compare the RC stochastic adaptive group-testing algorithm to a deterministic adaptive group-testing algorithm we refer to as SIGHT

(Sampling Inspired by Group Hyperedge Testing). SIGHT is a minor adaptation of traditional group testing strategies that have been designed to minimize the number of required tests to find all minimal defective sets [7, 9, 10], with small modifications to turn it into a sampling algorithm. Specifically, unlike [7, 9, 10], SIGHT (i) searches for only one of an unknown number $d$ of minimal defective $k$-sets, for $k_{min} \leq k \leq k_{max}$, and (ii) is tolerant of false negatives (*i.e.*, where a $k$-set that tests as non-defective actually contains at least one minimal defective $k$-set).

This paper is organized as follows: The SIGHT algorithm is described in Section 3.2.1, the RC algorithm in Section 3.2.2, and the test problem generator is described in Section 3.2.3. Experiments to compare the relative performance of the two algorithms for finding minimal defective $k$-sets are described in Section 3.2.4, with results presented in Section 3.3. We discuss the implications of our findings for some important applications in Section 3.4 and summarize our conclusions in Section 3.5. Although this study was originally motivated by the power systems problem described above, the implications are relevant to any hypergraph sampling problem.

## 3.2 Methods

### 3.2.1 SIGHT

#### Group Testing Inspiration

In the simple case of a single minimal defective edge in a graph, Triesch proposed a group testing halving procedure [7] that was later adapted by Johann [8] to find all $d$ defective edges (2-sets) in a simple graph in at most $d(\lceil \log_2 \frac{m}{d} \rceil + 7)$ tests, where $d$

is known. The algorithm presented by Johann for finding a single minimal defective edge in a subgraph $G' \subset G$ has subsequently been referred to as the TJ Procedure [9, 10] and is the primary inspiration for the $BinSearchSIGHT$ subroutine presented in Sec. 3.2.1. The TJ procedure was first used to find all $d$ defective edges, when $d$ is unknown, in simple graphs [9], and then extended to hypergraphs [10, 11]. SIGHT is a minor adaptation of the TJ procedure, modified in two ways: (i) it is designed to randomly sample a single defective $k$-set for $k_{min} \leq k \leq k_{max}$ (as opposed to assuming all defective sets of all sizes will be found), aborting if the minimal defective set has $k > k_{max}$; and (ii) it is designed to be robust to false negative tests.

**SIGHT Algorithm**

SIGHT takes as input the universal set $V$ of nodes, an initial subset size $a_0$, and the bounds $k_{min}$ and $k_{max}$ on the size of the defective $k$-sets one is searching for. It returns either a minimal defective $k$-set (for $k_{min} \leq k \leq k_{max}$) or the empty set (if the algorithm aborts). The only control parameter required by SIGHT is $a_0$, which establishes the size of the initial subset to be tested, where $k_{max} \leq a_0 < N$. The algorithm is illustrated by the Psuedocode in Algorithm 1 and is described below.

The first step of the SIGHT algorithm (Alg. 1: line 1) calls the subroutine $SAMPLE(V, a_0)$ (pseudocode not shown), which initializes a list $S$ to a uniform random sample, in random order, of $a_0$ unique elements from $V$. In the following description, the notation $S[i]$ is used to refer to the $i^{\text{th}}$ element of $S$; $S[i]$ is considered to be left of $S[j]$ for all $i < j$; and $S[:i]$ refers to the first $i$ elements of $S$.

A list $D$ is initialized to the empty list (Alg. 1: line 2). This list $D$ will be used

[1]The use of short-circuiting logic in the conditional statement prevents unnecessary tests.

**Algorithm 1** SIGHT$(V, a_0, k_{min}, k_{max})$

```
1: S ← SAMPLE(V, a_0)
2: D ← ∅
3: if ¬ isDefective(S) then
4:     return ∅
5: end if
6: while |D| < k_max do
7:     m ← BinSearchSIGHT(S, D)
8:     D ← D ++ S[m]
9:     if |D| ≥ k_min ∧ isDefective(D) then [1]
10:        D ← bottomUpSIGHT(D, k_min, k_max)
11:        return D
12:    end if
13:    S ← S[ : m − 1]
14: end while
15: return ∅
```

to accumulate known nodes from defective $k$-sets. In Alg. 1: line 3, the subroutine $isDefective(S)$ is called to test whether the list $S$ is a defective set (although not necessarily minimal). The algorithm $isDefective(S)$ is specific to the application problem, so is not shown here. It is *expected* to return TRUE if there is a minimal defective $k$-set in $S$ and FALSE otherwise; however, if false negatives can occur then $isDefective(S)$ may sometimes return FALSE even when $S$ contains a minimal defective set. If $isDefective(S)$ returns FALSE, the algorithm aborts and returns the empty set (Alg. 1: line 4). Otherwise, the algorithm calls the subroutine $BinSearchSIGHT(S, D)$ (Alg. 1: line 7).

The subroutine $BinSearchSIGHT(S, D)$ (Algorithm 2), following [8, 9, 10], uses a deterministic binary search to find the index of the leftmost element in $S$ that is the rightmost element of a defective $k$-set in $D +\!\!+ S$, where $+\!\!+$ denotes list concatenation. The subroutine is implemented like a classic binary search, in that it first assigns left

and right pointers $l$ and $r$ to the leftmost and rightmost elements of $S$, respectively (Alg. 2: lines 1-2), and determines a test index, $i$, half way between $l$ and $r$. But unlike a standard binary search that assumes that the values of the elements of $S$ are in sorted order and performs a test on element $S[i]$, here only the ordering of the indices matters (the actual elements in $S$ are deliberately in random order to prevent the algorithm from biasing towards sets that include elements with low values) and the test (Alg. 2: line 5) includes all elements of $S$ with indices $\leq i$ as well as all elements from the growing list $D$, to see whether $D + S[:r-i]$, is defective. If so, $r$ is reduced (Alg. 2: line 6); if not, $l$ is increased (Alg. 2: line 8). The process repeats until $l$ and $r$ converge to some index, which is returned.

**Algorithm 2** BinSearchSIGHT$(S, D)$

```
1: l ← 1
2: r ← |S|
3: while l < r do
4:     i ← ⌈(r−l)/2⌉
5:     if isDefective(D + S[:r − i]) then
6:         r ← r − i
7:     else
8:         l ← r − i + 1
9:     end if
10: end while
11: return r
```

After each call to the *BinSearchSIGHT* subroutine, SIGHT concatenates the found element $S[m]$ to $D$ (Alg. 1: line 8). If the accumulated set $D$ is found to be defective and is at least size $k_{min}$ (Alg. 1: line 9), then the algorithm has successfully found a list $D$ that contains a minimal defective $k$-set. However, in applications where false negatives can occur, $D$ may be non-minimal. To make the algorithm more robust to false negatives, it then calls the subroutine $bottomUpSIGHT(D, k_{min}, k_{max})$ (Alg.

1: line 10; pseudocode for *bottomUpSIGHT* not shown), which tests any subsets of $D$ of size $k_{min} \leq k \leq k_{max}$ that have not already been tested, to ensure that the defective subset returned is minimal (Alg. 1: line 11). If $D$ is not defective, then $S[m]$ and all elements to its right are removed from $S$ (Alg. 1: line 13) and the process is repeated until either $D$ tests as defective, or $|D|$ exceeds $k_{max}$ (Alg. 1: line 6), in which case the defective set being found is too large and the algorithm aborts (Alg. 1: line 15). Note that Alg. 1: lines 6, 9-12 comprise the minor differences between SIGHT and the previous approach to group testing on hypergraphs given in [10], on which SIGHT is based. Open source Matlab code for SIGHT is posted online [27].

**Computational Complexity of SIGHT**

The time complexity of $isDefective(S)$ is application-dependent, so computational complexity is here defined as a function of the number of tests required (*i.e.*, calls to $isDefective(S)$). Each call to $BinSearchSIGHT$ takes no more than $\lceil \log_2(a_0) \rceil$ tests to find an element of the defective set, and this operation is performed at most $k_{max}$ times before either a defective set is found or the algorithm aborts. Additionally, all subsets of $D$ larger than $k_{min}$ must be tested to ensure the defective set returned is minimal. The resulting upper-bound on the number of required tests by SIGHT is thus:

$$max(\#Tests|SIGHT) = k_{max}\lceil \log_2(a_0) \rceil + \sum_{j=k_{min}}^{k_{max}} \binom{k_{max}}{j} + 1 \tag{3.1}$$

In practice, this worst case is rarely realized. Since $a_0 < N$, each run of SIGHT requires $O(logN)$ tests.

However, not all tests are necessarily equal. In some applications, such as the power systems application described earlier, positive tests require much more computation time than negative tests. The required number of positive tests per each run of SIGHT is upper-bounded by:

$$max(\#PositiveTests|SIGHT) = k_{max}\lceil \log_2(a_0) \rceil \tag{3.2}$$

Note that this is explicitly a function of $k_{max}$.

### 3.2.2 Random Chemistry

#### Inspiration for RC

The basic idea for the RC algorithm was originally proposed by Kauffman [28] as a hypothetical method for identifying minimal auto-catalytic sets of interacting molecules from within a very large set of molecules. He suggested testing random half-sets until the products of auto-catalysis were detected in one subset, and repeating this halving process until a minimal auto-catalytic set was discovered (hence the moniker "Random Chemistry"). Although not previously presented as such, this is an adaptive stochastic group testing method.

Inspired by the RC idea, an RC algorithm was implemented for finding genetic interactions between single nucleotide polymorphims that predispose for disease [29]. Later, a version of RC was implemented for finding a small set of transmission line outages in electric power networks that trigger cascading power failure, requiring only $O(logN)$ tests per successful run [14].

### RC Algorithm

As in SIGHT, RC takes as input the set of nodes $V$ and returns either a single minimal defective $k$-set (for $k_{min} \leq k \leq k_{max}$), or the empty set (if the algorithm aborts). The bounds on $k$ are specified for RC exactly as they are in SIGHT. Pseudocode for RC, as implemented in this study, is provided in Algorithm 3.

**Algorithm 3** RC$(V, A, k_{min}, k_{max}, t_{max})$

```
1:  a_0 ← A(0)
2:  S ← SAMPLE(V, a_0)
3:  if ¬ isDefective(S) then
4:      return ∅
5:  end if
6:  for i ← 1, |A| do
7:      a_i ← A(i)
8:      t = 0
9:      flag ← false
10:     while t < t_max ∧ ¬flag do
11:         t ← t + 1
12:         S_new ← SAMPLE(S, a_i)
13:         if isDefective(S_new) then
14:             S ← S_new
15:             flag ← true
16:         end if
17:     end while
18:     if ¬flag then
19:         return ∅
20:     end if
21: end for
22: return bottomUpRC(S, k_min, k_max)
```

Like SIGHT, RC begins by drawing a random sample $S$ of elements from $V$ such that $|S| = a_0$ (Alg. 3: lines 1-2) and aborts, returning the empty set, if $S$ does not

contain a defective set (Alg. 3: lines 3-5). If an initial defective set $S$ of size $a_0$ is found, subset reduction proceeds stochastically according to the set size reduction scheme $A$ (loop starting at Alg. 3: line 6).

The sampling loop in RC (Alg. 3: lines 10-17) stochastically attempts to find a defective subset $S_{new}$ of size $a_i$, from set $S$ of size $a_{i-1}$. If no such subset is found after $t_{max}$ attempts, the algorithm aborts and returns the empty set (Alg. 3: lines 18-20). When a subset of size $a_{final}$ is found that causes a cascade, a bottom-up search is conducted (Alg. 3: line 22, which calls subroutine $bottomUpRC(S, k_{min}, k_{max})$, pseudocode not shown), testing all subsets of size $k$, for $k = k_{min}, \ldots, k_{max}$ (in random order for each $k$), returning either the first defective $k$-set found or the empty set, if no minimal defective set of size $\leq k_{max}$ exists in $S$. The subroutine $bottomUpRC$ differs slightly from $bottomUpSIGHT$, in that $bottomUpRC$ must test all subsets of $S$ until a defective subset is found or $k > k_{max}$. In SIGHT, some subsets of $S$ have already been tested during the binary search, so $bottomUpSIGHT$ only needs to test the subsets of $S$ that have not been previously tested. Open source Matlab code for RC, as implemented in this study, is posted online [30].

If there is exactly one minimal defective $k$-set, and the cost of each test is constant, then one can derive an optimal set reduction scheme $A$ for RC [31]. However, in many applications (such as the power systems application) it is not possible to analytically optimize $A$. In the work shown here, we use the subset reduction schedule of $a_i = a_{i-1}/c$, where $c = 2$ for $a_{i-1} > 20$ (binary splitting) and $c = 1.5$ for $a_{i-1} \leq 20$ (to increase success rate when groups get small), and specify $t_{max} = 20$, as proposed in [14, 16, 17, 18, 19]. The initial set size $a_0$ is tuned as a control parameter.

### Computational Complexity of RC

The RC algorithm requires a test of the initial subset, plus up to $t_{max}$ tests for each of the $|A| - 1$ reduction steps, plus additional tests for the bottom-up search of the set of size $a_{final}$. Thus, the maximum potential number of tests required by an RC run is:

$$max(\#tests|RC) = 1 + (|A| - 1)t_{max} + \sum_{k=k_{min}}^{k_{max}} \binom{a_{final}}{k} \tag{3.3}$$

In practice, (using the RC parameters specified in the experiments presented here) only a few tests ($\ll t_{max}$) are typically required during each stage of the subset reduction, and 2-sets are found most frequently (Section 3.3) even when $k_{max} > 2$, so the average number of tests required is much lower than this (Section 3.3). As long as $a_i = a_{i-1}/c$, for some constant(s) $c > 1$ (as implemented here), then $|A| \propto log(a_0)$, for some $a_0 < N$. Under these circumstances, each RC run requires $O(logN)$ tests.

In RC, each reduction step requires exactly one defective test, so the number of defective tests required by a successful RC run is constant. For a reduction scheme of length $|A|$ followed by a brute force search until one defective set is found, the maximum number of tests of defective sets required by an RC run is:

$$max(\#defectiveTests|RC) = |A| + 1 \tag{3.4}$$

Assuming set sizes in $A$ are fractionally reduced, 3.4 reduces to:

$$max(\#defectiveTests|RC) \leq \lceil \log(a_0) \rceil + 1 \tag{3.5}$$

Thus, the maximum number of defective tests is not a function of $k_{max}$, as it is in SIGHT (compare to Eq. (3.2)). Given that there is some unknown probability of aborting during any reduction step, the average number is less than this.

### 3.2.3 Tunable Test Problem Generator

We leverage DCSIMSEP as a convenient means of generating domain-independent test problems with large and variable (but unknown) numbers of minimal defective $k$-sets (for $k \geq 2$). In a power systems application, DCSIMSEP works as follows. Given an initial set of $k$ outages (e.g., transmission line failures), DCSIMSEP iteratively checks to see if some pre-defined threshold $T$ for system failure is exceeded. For modeling cascading power failures, $T$ might be some percent of the load that is shed [16] (as illustrated in Fig. 3.1) or some percent of components that become separated from the grid [14]. If the threshold $T$ is exceeded, the simulation terminates. If not, then the simulator calculates how the power generation would change, how power flow would be redistributed, and whether this would cause additional components to fail. These iterations continue until either the threshold $T$ is exceeded (a positive test), the system achieves a new equilibrium (a negative test), or some maximum amount of time has elapsed (also a negative test). DCSIMSEP has been shown to yield low, but non-zero, frequencies of false negatives [16].

For the experiments presented here, we run DCSIMSEP on a large synthetic power system, with a geographical topography based on the footprint of the 11-state western Unites States transmission system, which is included in the Electric Grid Test Case Repository [32]. The test case contains 10,000 buses (connection points, typically substations, through which generators provide power and loads draw power from the

network) and $N = 12,706$ branches (transmission lines and transformers). We limit our simulations to those in which $k$-sets comprise sets of extrinsically caused initiating branch outages, and we specify initial conditions that guarantee that the system is $N-1$ secure, using the approach described in [16]; *i.e.* there are no defective 1-sets.

For the purposes of using DCSIMSEP as a general test problem generator, rather than to model which $k$-sets cause cascading failures in power systems, we consider a set defective if DCSIMSEP simply exceeds a specified number of iterations ($T$). Due to the intractable number of unique $k$-sets, it is not computationally feasible to ascertain the exact numbers of minimal defective $k$-sets in these problems. However, because the number of minimal defective $k$-sets will decrease as $T$ increases, we are able to create test problems with varying numbers of minimal defective $k$-sets with $k \geq 2$.

Rather than using runtimes from DCSIMSEP as a measure of the variable computational costs of positive and negative tests, we assume that negative tests always require 1 unit of time to run and that all positive tests require some pre-specified constant units of time.

The resulting test problems and tunable P:N test cost ratios enable us to compare the relative efficiency of the SIGHT and RC sampling algorithms on arbitrary problems with different numbers of defective $k$-sets and different P:N test cost ratios.

### 3.2.4 Experiments

We report on results from 720,000 paired runs (*i.e.*, starting from the same random initial sets) of SIGHT and RC, with $k_{max} = 4$. Specifically, we report on 30,000 paired runs for each initial set size $a_0 \in \{16, 48, 80, 112, 144, 176\}$ and each test problem

threshold $T \in \{5, 15, 25, 35\}$. (We note that we also computed results for $k_{max} \in \{2, 3\}$, but only present results for $k_{max} = 4$, since the observed patterns were nearly identical for $k_{max} \in \{2, 3\}$ and $k_{max} = 4$ is the more general case.)

We record the number of positive and negative tests required for each run, where a run comprises a single call to SIGHT (Alg. 1) or RC (Alg. 3) on a given combination of $a_0$ and $T$ and either terminates successfully in a *find* (when a minimal defective $k$-set, with $k \in \{2, 3, 4\}$ is found) or is aborted. However, rather than reporting metrics *per run* of SIGHT or RC, we report metrics *per find*, as the latter amortizes in the cost of unsuccessful (aborted) runs by each sampling algorithm.

Since the resulting distributions of the number of positive tests and negative tests required per find of a defective $k$-set ($k \in \{2, 3, 4\}$) were non-Gaussian, we used non-parametric Mann Whitney $U$ tests to assess whether median values were significantly different between SIGHT and RC. Based on the median number of positive and negative tests per find on each test problem, we compute the expected computational costs of each algorithm at various P:N test cost ratios. Specifically, we assume that the cost of each negative test is 1 unit of time and the cost of each positive test is one of $\{1, 10, 50, 100\}$.

## 3.3 Results

RC and SIGHT found similar proportions of minimal defective 2-sets, 3-sets, and 4-sets over all initial set sizes $a_0$ and all thresholds $T$ tested (Fig. 3.2). The set size reduction approaches of both group-testing algorithms result in a bias toward finding smaller minimal defective $k$-sets, even though the total number of $k$-sets is known to

increase with increasing $k$ [19]. In most cases tested, both algorithms found 2-sets, 3-sets, and 4-sets in a ratio of roughly 6:3:1, although the proportion of higher order $k$-sets found is notably lower at $T = 5$, the case where there are the most defective $k$-sets, and at the smallest initial set size $a_0 = 16$, where the probability of including a higher order $k$-set is lowest.

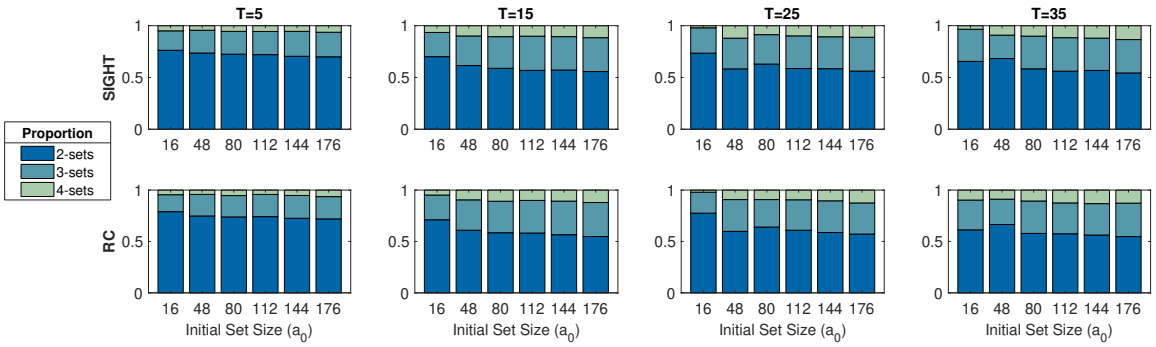



*Figure 3.2: The proportion of $k$-sets of size $k \in \{2, 3, 4\}$ found in 30,000 paired runs of SIGHT (top) and RC (bottom) with $k_{max} = 4$, for each $a_0 \in \{16, 48, 80, 112, 144, 176\}$ and each $T \in \{5, 15, 25, 35\}$.*

Despite this similarity in proportions, the two algorithms are not necessarily finding the same minimal defective sets, even though they start with identical random initial sets in each paired run (Fig. 3.3). This occurs because of the large number of minimal defective $k$-sets present in these test problems and the different set size reduction approaches taken by the two algorithms. Not surprisingly, the proportion of identical minimal defective sets found by SIGHT and RC drops as $a_0$ increases, because the number of minimal defective sets embedded in the initial set increases with increasing $a_0$. Similarly, for $a_0 \geq 48$, the proportion of identical sets found tends to increase with larger $T$, since increasing $T$ reduces the number of minimal defective sets in the system.

In nearly all circumstances tested, both algorithms require many more negative

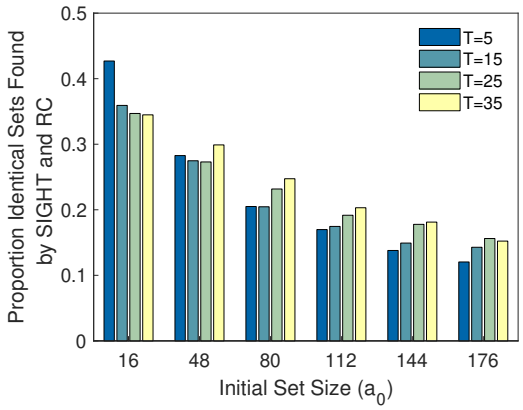



*Figure 3.3: The proportion of k-sets that were the same in 30,000 paired runs of SIGHT and RC with $k_{max} = 4$, for each $a_0 \in \{16, 48, 80, 112, 144, 176\}$ and each $T \in \{5, 15, 25, 35\}$.*

tests than positive tests per each successful find (the only exception being for SIGHT with $T = 5$ and $a_0 > 80$, due to the very large number of minimal defective sets present in the initial sets). For both algorithms, the median number of negative tests decreases rapidly as $a_0$ increases from 16 to 48 and then tends to plateau (Fig. 3.4, top row) while the number of positive tests required increases with increasing $a_0$ (Fig. 3.4, bottom row). SIGHT nearly always required significantly fewer negative tests than RC and exhibited less variability in the number of negative tests required per find (Fig. 3.5). The exception is when $a_0 = 16$ and $T \geq 15$, where the median number of negative tests was not significantly different; this occurs because the initial fail rate was over 99% and thus dominated the number of negative tests required by both algorithms. However, in all cases tested, SIGHT required significantly more positive tests than RC and exhibited greater variability in the number of positive tests per find (Fig. 3.6).

As expected, the initial failure rate (*i.e.*, the frequency with which the initial random set of size $a_0$ tests as non-defective) decreases monotonically both as the initial set size increases and as the number of defective sets decreases with increasing

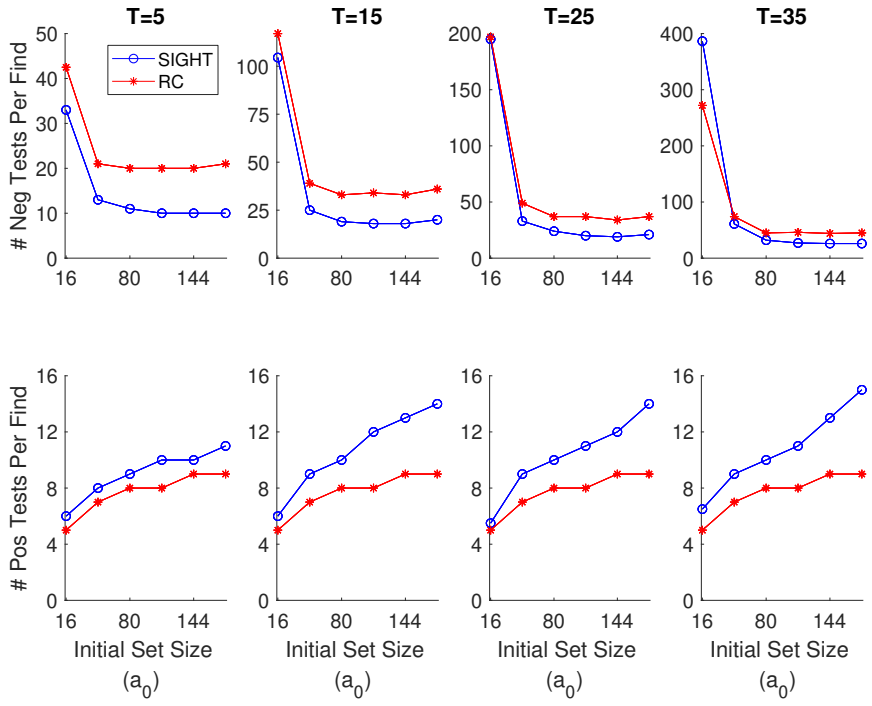



*Figure 3.4: Median number of positive and negative tests required per successful find of a minimal defective $k$-set of size $k \in \{2, 3, 4\}$. Note the differences in scaling of the $y$-axes.*

$T$ (Fig. 3.7). However, in runs where the initial set tests as defective, the failure rate due to subsequent aborts *increases* with increasing $a_0$ for both algorithms (Fig. 3.8; see Appendix for a proof of why this occurs).

Determining the optimal initial set size (*i.e.*, where expected runtime is lowest) is non-trivial. As the P:N test cost ratio increases, the optimal initial set size becomes smaller for both algorithms (Fig. 3.9, view top to bottom within each column). For a given P:N test cost ratio, the optimal $a_0$ becomes larger as the prevalence of minimal defective $k$-sets shrinks due to increasing $T$ (Fig. 3.9, view right to left within each row ).

For nearly all combinations of $a_0$ and $T$, the median number of total tests was significantly lower ($p < 0.005$) for SIGHT than for RC (Fig. 3.9, top row). (The

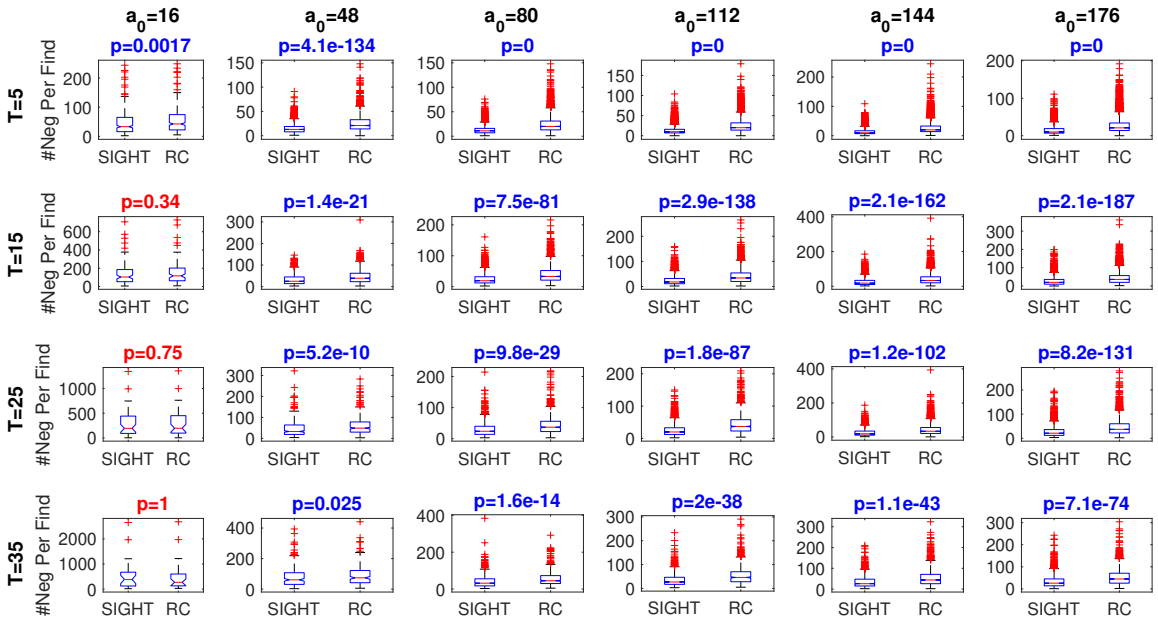



*Figure 3.5: Distributions of number of negative tests required per successful find of a minimal defective $k$-set of size $k \in \{2, 3, 4\}$. Blue p-values are statistically significant ($p < 0.05$). Note the differences in scaling of the y-axes.*

distributions were not significantly different when $T \geq 15$ and $a_0 = 16$ or when $T = 35$ and $a_0 = 48$, because the exceedingly high initial failure rates (Fig. 3.7) mean that negative tests on the paired initial sets dominate the required number of tests per find for both algorithms.) Thus, if positive and negative tests require the same computational cost, SIGHT is expected to be faster than RC. However, as the P:N test cost ratio increases, RC is expected to become faster than SIGHT (Fig. 3.9, view top to bottom) due to the higher number of positive tests required by SIGHT (Figs. 3.4, 3.6).

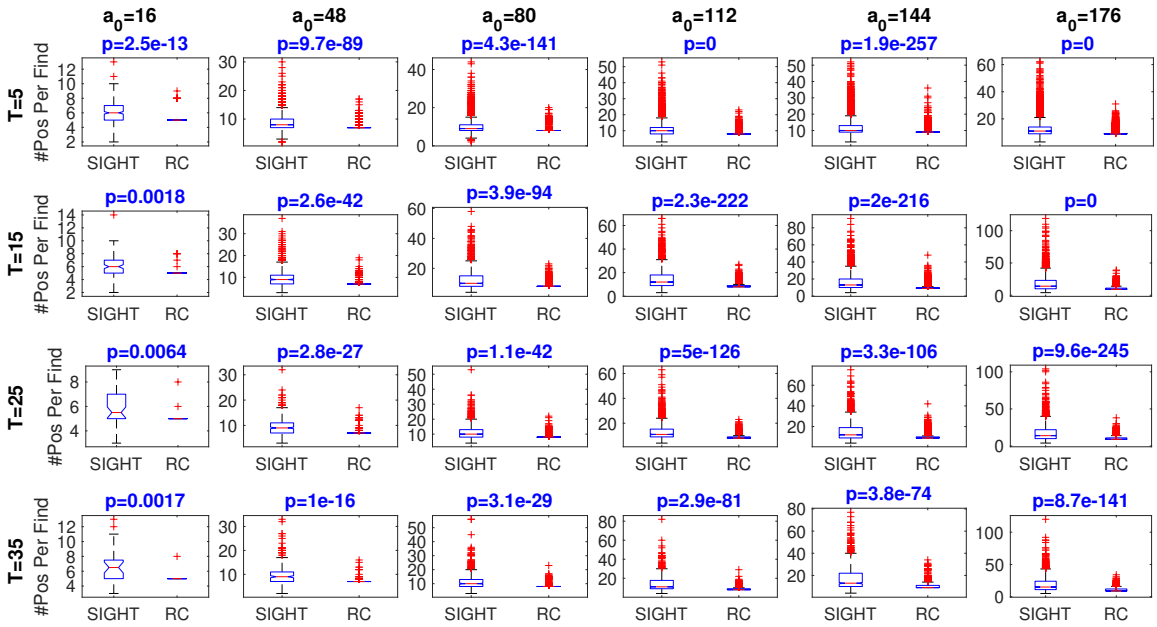



*Figure 3.6: Distributions of number of positive tests required per successful find of a minimal defective $k$-set of size $k \in \{2, 3, 4\}$. Blue $p$-values are statistically significant ($p < 0.05$). Note the differences in scaling of the $y$-axes.*

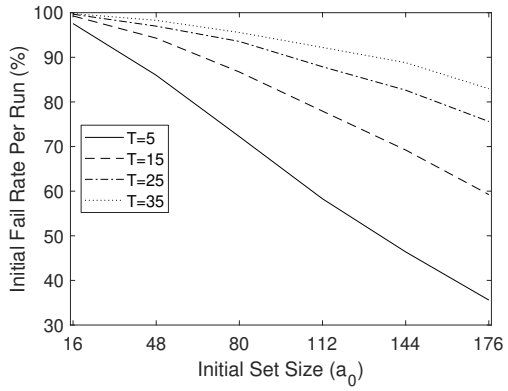



*Figure 3.7: The percent of runs that aborted due to the initial random set of size $a_0$ testing as non-defective. Since each paired run of SIGHT and RC started from the same initial random sets, this rate is identical for both algorithms.*

## 3.4 Discussion

In this work we compare deterministic (SIGHT) and stochastic (RC) adaptive group-testing algorithms on the problem of sampling minimal defective hyperedges (a.k.a.,

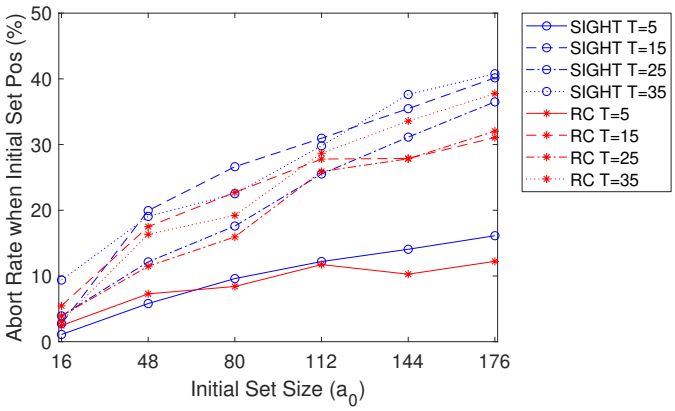



*Figure 3.8: The percent of runs in which the initial set of size $a_0$ was found defective but that were subsequently aborted.*

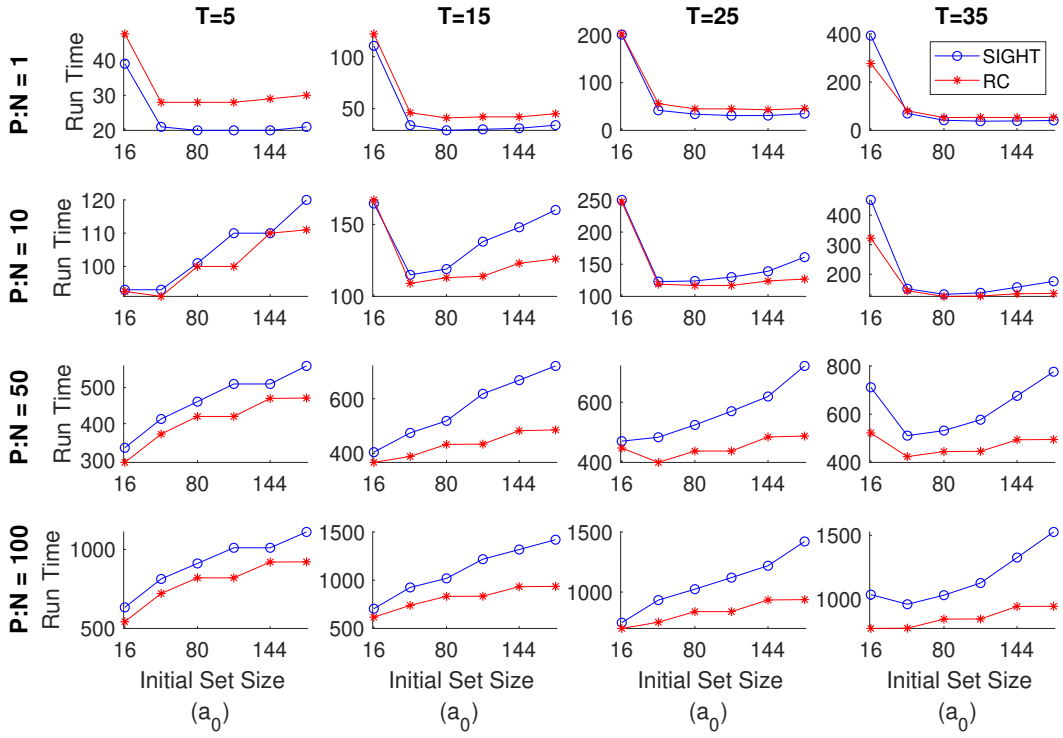



*Figure 3.9: Expected computational costs for SIGHT and RC per find at various P:N test cost ratios, for problems generated with the four thresholds $T$. The first row, where $P{:}N = 1$, also represents the median total tests required per find. Note the differences in scaling of the $y$-axes.*

$k$-sets), for $k \leq k_{max}$, in problems where there are intractably large numbers of defective $k$-sets ($k > 1$), false negatives can sometimes occur, and where the relative computational cost of testing defective sets and non-defective sets may differ.

Both algorithms are similarly effective in finding minimal 2-sets, 3-sets, and 4-sets, with a bias toward finding minimal defective $k$-sets with smaller $k$. While much of this bias is introduced in the random selection of the initial set (because smaller minimal defective $k$-sets are more likely to be included in the initial set), it is amplified during set reduction by different mechanisms in the two algorithms. In RC, with every subset reduction step there is a greater chance of preserving more smaller minimal defective $k$-sets than larger ones, since it is more likely that a larger minimal defective $k$-set will be disrupted during the sub-sampling procedure. Furthermore, the bottom-up search of the set of size $a_{final}$ at the end of RC, will always return one of the smallest contained minimal defective $k$-sets. In contrast, the binary search step in SIGHT searches for the leftmost element in $S$ that is the rightmost element of a defective $k$-set in $D + S$. Thus, the larger a minimal defective set is, the less likely it is to be selected by this procedure.

As a deterministic algorithm, SIGHT will always identify the same minimal defective $k$-set if starting from the same initial set; this is not the case for RC, where sets are stochastically reduced. In these experiments, the two algorithms found the same minimal defective $k$-sets less than half of the time in paired runs that started from identical initial sets.

The computational efficiency of both algorithms varies considerably with the initial set size $a_0$, so it is important to try to optimize this control parameter. It is obvious that, as $a_0$ increases, the chances of aborting a trial due to the initial set being

non-defective decreases (assuming false negatives are not predominant), as illustrated in Fig. 3.7. However, this type of failure is relatively computationally cheap, since it only requires a single test. In contrast, the percent of runs in which the initial set tests as defective but that later abort (due to the minimal defective set being larger than $k_{max}$) increases with increasing $a_0$ (Fig. 3.8; see Appendix for a proof of why this occurs). These latter types of failures can be quite computationally expensive, since they require many tests. The optimal $a_0$ finds the right balance between the relative frequencies of these two types of failures.

Our empirical results show that the optimal $a_0$, while similar for both algorithms on the same problem, is problem-specific. For example, the optimal $a_0$ is shown to be smaller for problems that have more defective sets (given the same P:N cost ratio). Similarly, the optimal $a_0$ is shown to be smaller for problems with higher P:N cost ratios (given the same number of defective sets).

Ultimately, the optimal $a_0$ is determined by the combined effects of (i) the steady increase in the number of positive tests required per find with increasing $a_0$ (Fig. 3.4, bottom row), (ii) the sharp drop in the number of negative tests required per find as $a_0$ increases from 16 to 48 (Fig. 3.4, top row), (iii) the (unknown) number of defective sets present in the system, and (iv) the P:N test cost ratio. The complex interplay of these factors is not possible to predict for real-world problems without simulating them. Thus, we recommend doing a parameter sweep to select an appropriate $a_0$, when applying either SIGHT or RC to a new problem.

We note that when RC was first presented for the power systems problem, we suggested that $a_0$ be selected to be large enough to achieve a low initial failure rate [14]. Accordingly, when we first applied RC to the large Western US Test case to assess

the risk of cascading failure, $a_0$ was chosen to be 320 in [18]. However, inspired by the new insights provided by the current study, we later found that reducing $a_0$ from 320 to 96 reduced run time per find, when using DCSIMSEP on this system, by nearly half when $k_{max} = 4$ (and by even more when $k_{max} < 4$), even though using the smaller $a_0$ increased the initial failure rate from 29% to 97%. Run times would be expected to be even more sensitive to $a_0$ if using a more computationally expensive AC simulator [21]. Assessing risk of cascading failure in a power system requires the identification of *many* minimal defective $k$-sets [16], and risk must be reassessed frequently (each time the load level changes) if it is to be mitigated [17]. Thus, choosing an appropriate $a_0$ will have a large impact on the computational practicality of using any group sampling approach to assess risk in power systems.

The reliability of tests also has a meaningful impact on the computational efficiency of both algorithms, but for different reasons. In SIGHT, a false negative test may cause an element of a minimal defective $k$-set to initially go undetected in the deterministic binary search portion of SIGHT, such that the defective set eventually found in the search is non-minimal. When this occurs, the element subsequently added to the growing set $D$ is not actually part of a minimal defective set. This is what necessitates the need for the computationally costly final search of $D$ to ensure that the defective set returned is minimal.

In contrast, while false negative tests can increase the required number of tests at a given set size for RC, the stochastic nature of subset selection in RC usually enables it to find a defective set despite the presence of false negatives. The bottom-up search of the final set of size $a_{final}$ is always required by RC, whether or not false negatives are present during the subset reduction steps.

The time complexity of successful runs of SIGHT and RC is $O(logN)$ in the number of tests required, for both algorithms. However, this does not mean that their average performances per find are the same. Overall, SIGHT requires fewer total tests than RC but more tests of defective sets, under nearly all circumstances tested. Thus, which algorithm is faster depends on the relative costs of testing defective *vs.* non-defective sets.

In the electric power system application that initially inspired this study, testing defective sets is generally much slower than testing non-defective sets, due to the high computational cost of simulating cascading failures, as shown using DCSIMSEP in Fig. 3.1. If using a more sophisticated AC simulator, this discrepancy will be even higher, since AC simulators are much slower [21]. Thus, we expect that RC will be faster than SIGHT for realistic problems in this power systems application, regardless of the particular power systems simulator used.

We note that the important question of whether existing DC or AC simulators are more appropriate for simulating cascading failures in power grids is outside the scope of, and indeed not even relevant to, this work. Our results are independent of what the minimal defective $k$-sets generated in these test problems represent.

Not all potential applications of these algorithms have higher costs for testing defective sets. For example, group testing has been proposed as a method for performing feature selection for classification tasks [33]. In this application, tests would determine whether a set of features achieves a specified level of performance on the classification task. "Defective" tests, in which a classifier exceeds the performance threshold, may actually be faster than "non-defective" tests, since after the threshold is reached classifier training can be aborted. Thus, it is expected that SIGHT would be

(potentially much) faster than RC in the feature selection application. Group testing has also been employed as a method for discovering synergistic reactions between drugs [34, 35, 36, 37], an application in which the cost of tests is constant; hence, since it requires fewer tests, SIGHT would be expected to outperform RC for this application.

## 3.5 Conclusions

This work compares, for the first time, deterministic (SIGHT) and stochastic (RC) adaptive group-testing methods for sampling minimal defective $k$-sets with variable $k \leq k_{max}$, where false negatives may be present, and where the computational costs of testing defective and non-defective sets may differ. We develop a test problem generator that enables us to vary the total number of minimal defective $k$-sets and we use a tunable parameter to control the relative computational costs of testing defective *vs.* non-defective sets.

Like RC, SIGHT is designed to sample from minimal defective sets of variable but bounded $k$, rather than identifying all minimal defective $k$-sets (as in prior works using deterministic group-testing on hypergraphs [10, 11]). In addition, the possible presence of false negatives has been largely ignored in the literature on adaptive group testing in graphs/hypergraphs [38, 39, 40], potentially causing them to fail. The slight modifications necessary to handle these conditions are all that distinguish SIGHT from its group testing ancestors [10, 11].

Both RC and SIGHT yielded similar distributions of minimal defective $k$-sets, when searching the same random initial defective sets, and exhibited similar sampling

bias toward finding minimal defective sets with lower $k$. The computational efficiency of both algorithms is sensitive to the selection of the initial set size $a_0$, which should be determined empirically since it is shown to be problem dependent. Despite their similarities, the computational properties of these two algorithms are shown to have important differences.

The existing group testing literature has treated all tests as having equal cost, and consequently the standard practice has been to try to develop deterministic algorithms that minimize the number of tests per defective set found [2]. SIGHT was, in fact, shown to require fewer total tests than RC. However, SIGHT was also shown to require more tests of defective sets than RC. Thus, which method is faster depends on the relative costs of testing defective *vs.* non-defective sets.

Our results indicate that the stochastic RC group testing algorithm is expected to outperform the deterministic SIGHT group testing algorithm in the power systems application that originally motivated this work, because positive tests are much more computationally costly than negative tests. In other applications, such as feature selection and drug discovery, where the cost of positive tests is less than or equal to the cost of negative tests, SIGHT is expected to out-perform RC.

## 3.6 Appendix

Empirically, the percent of runs that abort partway through a run (after the initial set tested as defective) increases with increasing $a_0$, for both algorithms (see Fig. 8 in the main text).

This occurs because the ratio of minimal defective sets of size $(k+1)$:$k$, in some

set $S$, increases as the size of the set increases, for all $k$. To prove this, consider a universal set $V$ of elements and the set $\Omega_k$, which contains all minimal defective $k$-sets, for a given $k$. Consider subset $S \subset V$ where the set sizes $|V| = N$ and $|S| = M$. The expected number of minimal defective $k$-sets in $S$ is $|\Omega_k^S| = \binom{M}{k}\binom{N}{k} \times |\Omega_k|$. Now, consider $T \subset V$ where $|T| = M + c$, for some positive constant $c$. Then, it suffices to show:

$$\frac{|\Omega_{k+1}^S|}{|\Omega_k^S|} < \frac{|\Omega_{k+1}^T|}{|\Omega_k^T|}$$

*Proof.*

$$\frac{\binom{M}{k+1}/\binom{N}{k+1}|\Omega_{k+1}|}{\binom{M}{k}/\binom{N}{k}|\Omega_k|} < \frac{\binom{M+c}{k+1}/\binom{N}{k+1}|\Omega_{k+1}|}{\binom{M+c}{k}/\binom{N}{k}|\Omega_k|}$$

$$\binom{M}{k+1}\binom{M+c}{k} < \binom{M}{k}\binom{M+c}{k+1}$$

$$\frac{M-k}{k+1}\binom{M}{k}\binom{M+c}{k} < \frac{M-k+c}{k+1}\binom{M}{k}\binom{M+c}{k}$$

$$M - k < M - k + c$$

□

Consequently, the number of expected minimal defective sets in $S$ will increase more slowly for $k \leq k_{max}$ (as in a successful run) than for $k > k_{max}$ (as in a run

that aborts because the found defective $k$-set is too large). The net effect is that, even though both RC and SIGHT are biased towards finding smaller $k$-sets, the rapid increase in the number of minimal defective sets that are too large, with increasing $a_0$, results in an increase in the proportion of runs that are aborted part way through a run in which the initial set was defective, for both SIGHT and RC.

# Bibliography


[1] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[2] Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 2000.

[3] Idan Yelin, Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagam Gandali, Tamar Hashimshony, et al. Evaluation of covid-19 rt-qpcr test in multi-sample pools. *medRxiv*, 2020.

[4] Cassidy Mentus, Martin Romeo, and Christian DiPaola. Analysis and applications of non-adaptive and adaptive group testing methods for covid-19. *medRxiv*, 2020.

[5] Martin Aigner. *Combinatorial search*. John Wiley & Sons, Inc., 1988.

[6] Peter Damaschke. A tight upper bound for group testing in graphs. *Discrete Applied Mathematics*, 48(2):101–109, 1994.

[7] Eberhard Triesch. A group testing problem for hypergraphs of bounded rank. *Discrete Applied Mathematics*, 66(2):185–188, 1996.

[8] Petra Johann. A group testing problem for graphs with several defective edges. *Discrete Applied Mathematics*, 117(1-3):99–108, 2002.

[9] Frank K Hwang. A competitive algorithm to find all defective edges in a graph. *Discrete Applied Mathematics*, 148(3):273–277, 2005.

[10] Ting Chen and Frank K Hwang. A competitive algorithm in searching for many edges in a hypergraph. *Discrete Applied Mathematics*, 155(4):566–571, 2007.

[11] Ting Chen. A revised algorithm for searching for all defective edges in a graph. *Discrete Applied Mathematics*, 159(18):2266–2268, 2011.

[12] Mohammad Reza Aghamohammadi and Ali Shahmohammadi. Intentional islanding using a new algorithm based on ant search mechanism. *International Journal of Electrical Power & Energy Systems*, 35(1):138–147, 2012.

[13] Paul DH Hines and Pooya Rezaei. *Smart Grid Handbook*, chapter Cascading Failures in Power Systems. John Wiley & Sons, 2016.

[14] Margaret J Eppstein and Paul DH Hines. A "random chemistry" algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

[15] Pooya Rezaei, Paul DH Hines, and Margaret J Eppstein. Estimating cascading failure risk: Comparing monte carlo sampling and random chemistry. In *2014 IEEE PES General Meeting | Conference & Exposition*, pages 1–5. IEEE, 2014.

[16] Pooya Rezaei, Paul DH Hines, and Margaret J Eppstein. Estimating cascading failure risk with random chemistry. *IEEE Transactions on Power Systems*, 30(5):2726–2735, 2014.

[17] Pooya Rezaei, Margaret J Eppstein, and Paul DH Hines. Rapid assessment, visualization, and mitigation of cascading failure risk in power systems. In *2015 48th Hawaii International Conference on System Sciences*, pages 2748–2758. IEEE, 2015.

[18] Laurence A Clarfeld, Margaret J Eppstein, Paul DH Hines, and Eric M Hernandez. Assessing risk from cascading blackouts given correlated component failures. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7. IEEE, 2018.

[19] Laurence A Clarfeld, Paul DH Hines, Eric Hernandez, and Margaret J Eppstein. Risk of cascading blackouts given correlated component outages. *IEEE Transactions on Network Science and Engineering*, 2019.

[20] Hassan Haes Alhelou, Mohamad Esmail Hamedani-Golshan, Takawira Cuthbert Njenda, and Pierluigi Siano. A survey on power system blackout and cascading events: Research motivations and challenges. *Energies*, 12(4):682, 2019.

[21] Wenyun Ju. *Modeling, Simulation, and Analysis of Cascading Outages in Power Systems*. PhD thesis, University of Tennessee, 12 2018.

[22] Yang Yang, Takashi Nishikawa, and Adilson E Motter. Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365):eaan3184, 2017.

[23] Hengdao Guo, Ciyan Zheng, Herbert Ho-Ching Iu, and Tyrone Fernando. A critical review of cascading failure analysis and modeling of power system. *Renewable and Sustainable Energy Reviews*, 80:9–22, 2017.

[24] Harsha Nagarajan, Russell Bent, Pascal Van Hentenryck, Scott Backhaus, and Emre Yamangil. Resilient transmission grid design: Ac relaxation vs. dc approximation. *arXiv preprint arXiv:1703.05893*, 2017.

[25] Hale Cetinay, Saleh Soltan, Fernando A Kuipers, Gil Zussman, and Piet Van Mieghem. Comparing the effects of failures in power grids under the ac and dc power flow models. *IEEE Transactions on Network Science and Engineering*, 5(4):301–312, 2017.

[26] Russell Bent, Carleton Coffrin, Rodrigo RE Gumucio, and Pascal Van Hentenryck. Transmission network expansion planning: Bridging the gap between ac heuristics and dc approximations. In *2014 Power Systems Computation Conference*, pages 1–8. IEEE, 2014.

[27] Laurence A Clarfeld. Sampling inspired by group hyperedge testing (sight). https://www.mathworks.com/matlabcentral/fileexchange/72718-sampling-inspired-by-group-hyperedge-testing-sight, 2019-09-11.

[28] Stuart Kauffman. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford university press, 1996.

[29] Margaret J Eppstein, Joshua L Payne, Bill C White, and Jason H Moore. Genomic

mining for complex disease traits with "random chemistry". *Genetic Programming and Evolvable Machines*, 8(4):395–411, 2007.

[30] Laurence A Clarfeld. Random chemistry (rc). https://www.mathworks.com/matlabcentral/fileexchange/72720-random-chemistry-rc, 2019-09-11.

[31] Jeffrey S Buzas and Gregory S Warrington. Optimized random chemistry. *arXiv preprint arXiv:1302.2895*, 2013.

[32] Adam B Birchfield, Ti Xu, Kathleen M Gegner, Komal S Shetye, and Thomas J Overbye. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on Power Systems*, 32(4):3258–3265, 2017.

[33] Yingbo Zhou, Utkarsh Porwal, Ce Zhang, Hung Q Ngo, XuanLong Nguyen, Christopher Ré, and Venu Govindaraju. Parallel feature selection inspired by group testing. In *Advances in Neural Information Processing Systems*, pages 3554–3562, 2014.

[34] Katja S Remlinger, Jacqueline M Hughes-Oliver, S Stanley Young, and Raymond L Lam. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, 48(1):133–143, 2006.

[35] Bryan Severyn, Robert A Liehr, Alex Wolicki, Kevin H Nguyen, Edward M Hudak, Marc Ferrer, Jeremy S Caldwell, Jeffrey D Hermes, Jing Li, and Matthew Tudor. Parsimonious discovery of synergistic drug combinations. *ACS Chemical Biology*, 6(12):1391–1398, 2011.

[36] Jacqueline M Hughes-Oliver. Pooling experiments for blood screening and drug discovery. In *Screening*, pages 48–68. Springer, 2006.

[37] Alexis A Borisy, Peter J Elliott, Nicole W Hurst, Margaret S Lee, Joseph Lehár, E Roydon Price, George Serbedzija, Grant R Zimmermann, Michael A Foley, Brent R Stockwell, et al. Systematic discovery of multicomponent therapeutics. *Proceedings of the National Academy of Sciences*, 100(13):7977–7982, 2003.

[38] Kyle Li, Doina Precup, and Theodore J Perkins. Pooled screening for synergistic interactions subject to blocking and noise. *PloS One*, 9(1):e85864, 2014.

[39] Huilan Chang, Hong-Bin Chen, and Hung-Lin Fu. Identification and classification problems on pooling designs for inhibitor models. *Journal of Computational Biology*, 17(7):927–941, 2010.

[40] Fei-huang Chang, Huilan Chang, and Frank K Hwang. Pooling designs for clone library screening in the inhibitor complex model. *Journal of Combinatorial Optimization*, 22(2):145–152, 2011.

# Chapter 4

# A General Model of Conversational Dynamics and an Example Application in Serious Illness Communication

## Abstract

Conversation has been a primary means for the exchange of information since ancient times. Understanding patterns of information flow in conversations is a critical step in assessing and improving communication quality. In this paper, we describe **CO**nversational **DY**namics **M**odel (CODYM) analysis, a novel approach for studying patterns of information flow in conversations. CODYMs are Markov Models that capture sequential dependencies in the lengths of speaker turns. The proposed method is automated and scalable, and preserves the privacy of the conversational participants. The primary function of CODYM analysis is to quantify and visualize patterns of information flow, concisely summarized over sequential turns from one or more

conversations. Our approach is general and complements existing methods, providing a new tool for use in the analysis of any type of conversation. As an important first application, we demonstrate the model on transcribed conversations between palliative care clinicians and seriously ill patients. These conversations are dynamic and complex, taking place amidst heavy emotions, and include difficult topics such as end-of-life preferences and patient values. We perform a versatile set of CODYM analyses that (a) establish the validity of the model by confirming known patterns of conversational turn-taking and word usage, (b) identify normative patterns of information flow in serious illness conversations, and (c) show how these patterns vary across narrative time and differ under expressions of anger, fear and sadness. Potential applications of CODYMs range from assessment and training of effective healthcare communication to comparing conversational dynamics across language and culture, with the prospect of identifying universal similarities and unique "fingerprints" of information flow.

## 4.1 Introduction

Conversation is a fundamental form of human communication. Conversations are highly complex phenomena [1], but use simple rules to maintain discourse [2]. Humans have an innate ability to learn spoken language from infancy, yet despite the importance of conversations in our daily lives, achieving effective communication through conversations can be difficult [3, 4]. Developing a better understanding of information flow in different conversational contexts can help guide efforts to improve conversation quality.

Conversation analysis (CA) became established as a discipline of study beginning in the late 1970's with the seminal work of Harvey Sacks and others, such as in their formative 1978 paper [2], in which they present a framework for the process of conversation. Sacks described conversation as being highly structured around turn-taking, with participants being able to fluidly transition between turns without turns overlapping. This fundamental property of conversation has subsequently been

observed and measured across languages and cultures, revealing itself to be a universal trait [5]. Sacks theorized that this discourse is maintained by a set of rules, or norms, that are followed by participants and govern when the speaking floor is relinquished and which participant may speak next. Perhaps the most important aspect of his framework is the heavy focus on the sequential nature of conversation and the dependence of each speaker turn on the turns that came before it [2].

The traditional conversation analytic approach to understanding sequence is to use meticulous transcriptions of recorded conversations in order to study a pre-specified conversational phenomenon and understand its normative patterns [6, 7]. For example, in one early, influential study, Schegloff examined the opening sequence of turns in 500 telephone calls and attempted to explain the patterns he observed [8]. This approach has been widely adopted and applied in diverse contexts, resulting in a vast body of work comprising thousands of research papers. While the value and importance of this inherently qualitative approach remains relevant today [6, 7, 9], quantitative methods have gained increasing popularity in CA (e.g., [10, 11]). In particular, Markov Models (MMs) inherently model sequential events [12], and so have been widely applied in CA.

In a MM, the likelihood of a given event occurring is determined by the current state of the system, and when an event does occur it causes the system to transition to a new state [12]. The "order" of a MM defines the number of previous events that are recorded in each state (i.e., the length of the "memory" in the model). In most CA applications of MMs, state transitions are defined to take place between some constant, fixed intervals of time. Examples of these include $1^{st}$-order MMs that were used to classify dialog scenarios in conversations based on speech/silence states

[13], and for identifying conversational structure within non-verbal states, such as gaze patterns in four-person conversations [14], and 2nd-order MMs used to study the effects of conversational speech/silence patterns on communication systems [15]. In contrast, sequences of speaker order in four-person conversations were used to predict who the next speaker would be, using MMs up to order 5 [16]; 2nd-order MMs proved significantly better than 1st-order MMs for this task, but little further improvement was gained by moving to higher order models on this data set, and a simple context-sensitive model based on speaker roles was shown to out-perform the MMs.

While the examples above use MMs as a tool for making predictions or classifications regarding conversations, another approach to understanding the structure of information flow involves classifying units of conversation by their functional roles (e.g., [17, 18]). Once these functional roles have been defined, 1st-order Markov models have been used to understand the sequence of these functions in conversation [19, 20, 21]. Influence modeling is yet another Markov-based method, where individual Markov chains for each speaker are coupled together to understand how speakers interact, including understanding which speakers are most influential [22, 23, 24] and the functional role of each speaker [25].

Visualizing data to aid interpretation has also grown in popularity [26], and a number of methods have been proposed for visualizing conversational dynamics [27, 28, 29]. Good visualizations allow complex data to be displayed in a more easily digestible format, allowing readers to better recognize and understand patterns that may not otherwise be apparent [30]. One popular tool for visualizing conversational dynamics is Discursis [29], which uses conceptual recurrence plots for unsupervised

identification and visualization of shared content between speaker turns in the analysis of conversational discourse. This method has been used in a variety of contexts, including the study of healthcare conversations [31, 32, 33, 34, 35, 36]. Discursis visualizations portray the lengths of each speaker turn throughout an individual conversation, along with the amount of overlap in content in these turns. Thus, the size and complexity of Discursis visualizations vary for each conversation, and access to full transcriptions are necessary to create these very detailed visualizations of individual conversations.

In this paper, we describe the **CO**nversational **DY**namics **M**odel (CODYM), a novel approach for analyzing and visualizing information flow across sequences of turns in one or many conversations, using $2^{\text{nd}}$-order and $3^{\text{rd}}$-order MMs of discretized turn lengths. CODYMs are based on the assumption that the length of a speaker turn is a simple proxy for the capacity of information conveyed in the turn, and that the amount of information conveyed during a given turn is influenced by the amount of information conveyed in previous speaker turns. The proposed method is scalable and can be fully automated. Since CODYMs do not rely on knowledge of the specific content of conversations, they do not compromise the privacy of the conversational participants. However, CODYMs can also be contextualized to study information flow patterns surrounding specific topics of interest, if such are known. In contrast to previous applications of MMs to conversation analysis, the primary function of CODYM analysis is to quantitatively summarize and visualize information flow patterns throughout one or more conversation(s), rather than to make predictions or classifications. Our approach is general and provides a new tool for use in the analysis of any type of conversation.

In healthcare communication, and especially in serious illness communication, the quality of doctor-patient conversations can have profound tangible impacts [37, 38, 39]. Promoting high-quality communication in serious illness healthcare is considered a national priority [40, 41], and there is an increasing recognition that automated methods for analyzing clinical conversations, such as turn-taking analysis, could provide useful feedback and insights for improving communication between clinicians and patients [39, 42]. Thus, as an important first application, we apply CODYM analysis to a corpus of 355 transcribed conversations between palliative care clinicians and seriously ill patients, recorded as part of the Palliative Care Communication Research Initiative (PCCRI) [43]. These conversations are dynamic, complex phenomena that take place amidst heavy emotions such as anger, fear, and sadness [44]. They include difficult topics such as end-of-life preferences and values, all while patients endure suffering from the symptoms of their illness. We seek to answer the following questions. What are the normative information flow patterns in serious illness conversations? How do those patterns change during the course of a conversation? Do certain words or topics tend to appear more in one information sharing pattern than another? How does the expression of distressing emotion impact information flow? We show that CODYM analysis provides a quantitative approach, with an intuitive interpretation, that helps to answer these questions.

The remaining sections of this manuscript are organized as follows. We first describe the methods involved in CODYM analysis, and the PCCRI corpus of palliative care conversations. We then present results of applying CODYM analysis to the PCCRI corpus in various ways that demonstrate the model's versatility, followed by a discussion of the significance of our findings and their relation to the current literature.

Finally, we close with some general conclusions and ideas for future work.

## 4.2 Methods

### 4.2.1 Conversational dynamics model

Due to the sequential nature of conversations [2], we expect that the amount of information conveyed in each turn influences the amount of information conveyed in subsequent turns. Here, we propose using the number of words in a speaker turn as a simple proxy for the capacity of information that the turn can convey. We define a **CO**nversational **DY**namics **M**odel (CODYM) to be a Markov Model (MM), where events are speaker turns of a given length and states comprise the lengths of some number (defined by the order of the model) of previous turns. CODYMs thus model the sequential patterns in turn lengths.

Any MM requires a discrete state-space, so turn lengths in a CODYM are discretized into a finite number of bins. Although, in principle, turn lengths can be discretized into any number of bins, the most appropriate number of bins will depend, in part, on the size of the data set. In the PCCRI corpus used here, turn lengths follow a heavy-tailed distribution, with a median turn length of 7 (Fig. 4.9). Here, we binarize turn lengths into short (`S`) turns and long (`L`) turns, with short turns defined as those with 1-7 words and long turns as those with 8 or more words. Using the median turn length as the maximum length of short turns (a) creates a relatively balanced data set with 53,751 short turns and 47,812 long turns, and (b) maximizes the Shannon

entropy (a measure of information content) for the distribution of states in a 3rd-order CODYM (Fig. 4.10). In preliminary experimentation with the PCCRI corpus, we found that discretizing turn lengths into ternary bins (Short/Medium/Long) was problematic because it both (a) created sample size issues by reducing the number of turns associated with each transition, and (b) resulted in more complex models that were difficult to interpret.

Note that the number of states in an $N^{\text{th}}$-order CODYM of binarized turn lengths has $2^N$ states and $2^{N+1}$ transitions. For example, a 2nd-order CODYM has 4 states and 8 transitions, and a 3rd-order CODYM has 8 states and 16 transitions, as illustrated in Fig. 4.1, where states are represented as nodes and transitions are represented as directed edges in a network. We use 2nd- and 3rd-order CODYMs in our analyses for the reasons explained below.

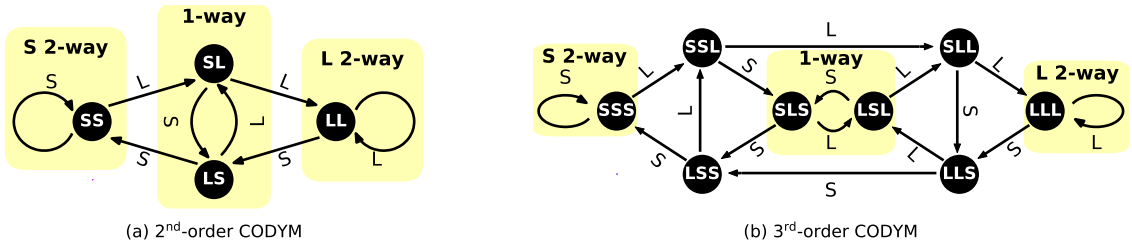



*Figure 4.1:* ***Network representations of CODYMs.*** *Network depictions of (a) 2nd-order and (b) 3rd-order CODYMs, where turn lengths are binarized as short (S) or long (L). Nodes (black circles) represent states that are defined by the lengths of the 2 or 3 previous turns, respectively; edges (arrows) represent transitions between states and are labeled with the length of the turn on that transition. The areas highlighted in yellow represent important sub-networks we refer to as short two-way information exchanges (labeled "S 2-way"), one-way information exchanges (labeled "1-way"), and long two-way information exchanges (labeled "L 2-way").*

We use 3rd-order CODYMs for analyzing normative patterns in serious illness conversations, and for examining how these patterns change both temporally and when distressing emotions are expressed. We selected 3rd-order CODYMs for this because:

(a) for dyadic conversations, such as between a patient and a clinician, having memory of 3 previous turns can represent a complete back-and-forth exchange between the 2 sides; (b) higher-order models become difficult to interpret; (c) large state spaces result in fewer observations of each state/transition, potentially resulting in small sample sizes that may preclude accurate characterization of normative patterns (e.g., the median numbers of occurrences of the different states for each conversation in the PCCRI corpus are 57, 28, 13, 6, and 3, for CODYMs of orders 1-5, respectively); and (d) we found that features taken from 3$^{\text{rd}}$-order CODYMs had greater predictive power over those from 1$^{\text{st}}$- or 2$^{\text{nd}}$-order CODYMs on an emotion-based classification task (described in more detail later), as shown in Table 4.2. On the other hand, a 2$^{\text{nd}}$-order CODYM has half the number of transitions as a 3$^{\text{rd}}$-order model (Fig. 4.1). Thus, when studying patterns of word associations with transitions, we use 2$^{\text{nd}}$-order CODYMs to increase the number of words associated with each of the transitions.

Different sub-networks of a CODYM can be interpreted as distinct regimes of information flow (highlighted in yellow in Fig. 4.1). For example, we refer to the center loop ($\mathsf{SL} \overset{S}{\rightarrow} \mathsf{LS} \overset{L}{\rightarrow} \mathsf{SL}$, in a 2$^{\text{nd}}$-order CODYM, and $\mathsf{SLS} \overset{L}{\rightarrow} \mathsf{LSL} \overset{S}{\rightarrow} \mathsf{SLS}$, in a 3$^{\text{rd}}$-order CODYM) as "one-way information exchanges", because alternation between $\mathsf{S}$ and $\mathsf{L}$ turns in a dyadic conversation implies that one party is conveying most of the information. We refer to the leftmost self-loop ($\mathsf{SS} \overset{S}{\rightarrow} \mathsf{SS}$, in a 2$^{\text{nd}}$-order CODYM, and $\mathsf{SSS} \overset{S}{\rightarrow} \mathsf{SSS}$, in a 3$^{\text{rd}}$-order CODYM) as "short two-way information exchanges". Conversely, we refer to the rightmost self-loop ($\mathsf{LL} \overset{L}{\rightarrow} \mathsf{LL}$, in a 2$^{\text{nd}}$-order CODYM, and $\mathsf{LLL} \overset{L}{\rightarrow} \mathsf{LLL}$, in a 3$^{\text{rd}}$-order CODYM) as "long two-way information exchanges". As we will show, each of these information flow regimes can be associated with different functions in conversations.

### 4.2.2 Observing patterns of information flow

In most MMs, the weights on all outgoing edges of each given node sum to 1.0, where each edge weight represents the probability that the node is left *via* that edge. This is appropriate when MMs are used as generative models or to make predictions of future states. However, with CODYMs our primary intent is to study patterns of information flow through all states and transitions in an existing corpus of dialog, not to generate simulated sequences of short and long speaker turns or to predict subsequent turn lengths. Thus, in a CODYM, the weights on nodes and edges represent their respective percentage frequencies of occurrence over all turns being analyzed. Consequently, the sum of all edge weights (transition frequencies) in an entire CODYM is 100% and the sum of all node weights (state frequencies) is also 100%.

We "populate" a CODYM by computing observed frequencies of each state/transition across a specified set of speaker turns. This set may comprise all turns in an entire corpus of conversations, all turns for a given speaker, all turns within individual conversations, or some other subset of turns that satisfy some pre-specified condition, depending on the question being addressed. When different CODYMs are populated separately for a number of conversations, they can be visualized as a single CODYM populated with the mean weights for each of states and transitions, averaged over all of the CODYMs populated from individual conversations. A CODYM of mean frequencies can be interpreted as a representation of the overall "normative" pattern of information flow in the set of conversations under study, assuming the distributions of frequencies are uni-modal for each state and transition. Alternatively, when CODYMs are populated from a subset of turns in the corpus, turns from multiple conversations

may be pooled to increase sample sizes in the states/transitions; in this case, we compute a single set of CODYM frequencies for the pooled data.

Where possible, we seek to determine whether a CODYM computed from observed data exhibits state and/or transition frequency distributions that differ from what would be expected if the given conversational feature of interest was independent of prior sequences of turn lengths (the null hypothesis). How we determine "expected" frequencies, and how we compare observed frequencies to them, depends on the particular experiment. For example, we compare two sets of observed CODYM frequency distributions, computed from turns with or without a given feature present, using 2-sample Kolmogorov-Smirnov tests. However, in most cases, we generate expected frequencies from appropriate null CODYMs, which are populated from data that have been randomly sampled (or randomly reorganized) in such a way as to disrupt any possible association between the conversational feature of interest and previous sequences of turn lengths, while preserving other salient characteristics of the data, such as sample size and the number of long and short turns used by patients and clinicians. When comparing to null CODYMs created by random sampling, we derive empirical probability distributions for null models computed from randomly sampled data by generating 1000 random copies of each, using Monte Carlo (MC) simulations. (Prior experimentation had shown that probability distributions were quite stable when created with 1000 MC simulations). When we visualize these null CODYMs, we display mean state and transition frequencies, averaged over all 1000 copies. If observed state and/or transition frequencies are outside of the empirically derived 95% confidence interval from the distribution of 1000 corresponding null models, the difference is considered to be statistically significant at the $p < 0.05$ level. To minimize

confusion, we provide the specifics of how we determined expected values for each type of experiment in the Results.

State space diagrams of CODYMs are displayed in one of two formats, based on what we believe most effectively conveys the results we are trying to highlight. In some cases, we display transition frequencies directly ($\%Observed$ or $\%Expected$), while in other cases we display mean observed transition frequencies minus the mean expected frequencies of the corresponding null models ($\Delta frequency = \%Observed - \%Expected$). In the former, all state weights sum to 100%, all transition weights sum to 100%, and the color bars represent only positive percentages. In the latter, transition weights sum to 0%, negative values indicate the degree to which the feature of interest is under-represented on a given transition, and positive values indicate the degree to which it is over-represented, relative to the corresponding null model. We use different colormaps to help distinguish these two visualization techniques. In both visualization formats, the thickness and color of transition arrows indicate the magnitude of the corresponding transition weights, and nodes are sized according to the frequency of their respective states. Where relevant, state and transition frequencies that are statistically significantly different from the corresponding null model are indicated by coloring states black (*vs.* gray) and drawing transition arrows with solid lines (*vs.* dashed lines).

To assess how normative patterns in information flow may change over the course of a conversation, we divide the turns of each conversation in the corpus into sequential deciles of words (ten bins of narrative time, as in [45]), stratified by patient and clinician turns. Note that different conversations have different numbers of turns, so the number of turns per bin varies by conversation. Individually, conversations average

only about 28 turns per decile, split approximately equally between patient turns and clinician turns. Since 14 turns are inadequate to robustly determine frequencies on 16 transitions, we pool the data by summing the number of patient and clinician turns per decile over all conversations. We then compute the frequencies of the pooled turns on each of the 16 transitions in 3$^{\text{rd}}$-order CODYMs, one per decile.

### 4.2.3 Contextualization of CODYMs

A CODYM is based exclusively on turn lengths, so is thus independent of what is actually said or expressed during those turns. However, a CODYM can be used to examine the information flow patterns involving different words, topics, expressions of emotion, or more generally, "contextual events".

We use two complementary approaches to study how word usage varies between conversational regimes. In the first approach, we use an unsupervised cluster analysis to determine which commonly used words have similar CODYM patterns of information sharing. Of the 14,848 unique words that appear in the PCCRI corpus, we only consider those that appear at least 100 times and for which the absolute value of $\%Observed - \%Expected$ is greater than 10; expected values are the observed frequencies in a 2$^{\text{nd}}$-order CODYM populated from all words over all turns in the corpus (Fig. 4.2). We use this method for prefiltering words because: (a) requiring a minimum number of occurrences ensures that a word is frequent enough in the corpus to asses its typical usage; and (b) by considering only words whose information flow patterns differ substantially from overall word frequencies, we focus on words that have specialized usages with respect to the normative information flow pattern.

The resulting list of 114 words was clustered using a standard K-Means clustering

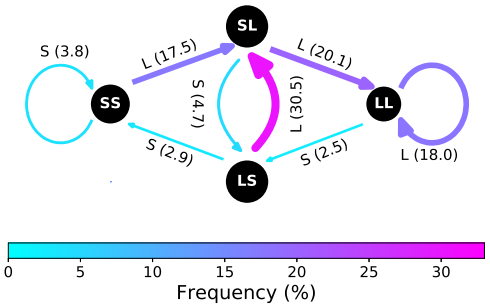



*Figure 4.2:* ***CODYM of word usage in PCCRI corpus.*** *Frequencies of occurrence of all words in the PCCRI corpus on transitions of a $2^{nd}$-order CODYM. Transition labels indicate the length of the turn on that transition, parenthetically followed by the percentage of word occurrence on that transition. Edge thickness and color indicate %Observed for each transition, and node diameter indicates %Observed for each state.*

algorithm (implemented with Python's scikit-learn package, with default settings [46]), using the 8 transition frequencies from $2^{\text{nd}}$-order CODYMs (computed for each of the 114 words) as the input features for clustering. Note that each of the 114 words are weighted equally during this clustering, even though some words occurred much more frequently than others. After qualitatively assessing cluster membership when using between three and ten clusters, we determined that using six clusters resulted in the most logically cohesive word groupings.

In the second approach, we populate CODYMs using frequencies of occurrence of terms from pre-defined lists. In this work, we focus on two such lists of interest in serious illness conversations: hedging terms and treatment terms.

There are many definitions of uncertainty terms and subtypes used in clinical conversations (e.g., [47, 48, 49, 50]). Here we focus on a list of "hedging" terms, recently developed for use in natural language processing of serious illness conversations, which was found to be the most prevalent subtype of uncertainty expressed in the PCCRI corpus [51]. According to this definition, hedging occurs in a conversation when the

speaker makes a word choice to suggest something rather than state it as fact, such as to say "I expect you will feel better" instead of "You will feel better". Using the word *expect* injects some uncertainty into the statement. Some examples of hedging terms include expect, hint, imply, perceive, presume, speculate, suspect, and think. The complete list of 35 hedging terms used in this study, and the frequency of occurrence of each term in the PCCRI corpus, is shown in Table 4.3. See [51] for a detailed description of how this list was compiled.

Treatment terms are those used to discuss medical care provided to patients. A list of 51 treatment words occurring at least 100 times each in the PCCRI corpus was previously prepared, and it was shown that these words vary temporally throughout the PCCRI conversations, being most prevalent in deciles 4 and 5 [45]. Some examples of treatment terms include antibiotics, ICU, chemo, fentanyl, machine, milligram, procedure, radiation, and ventilator. The complete list of the treatment terms used in this study (from [45]), and the frequency of occurrence of each term in the PCCRI corpus, are shown in Table 4.4.

Note that specific hedging terms or treatment terms that are more frequent in the overall corpus have a greater influence on resulting state and transition frequencies, since CODYM frequencies were computed over all occurrences of all terms in a given list. Thus, to assess robustness of the observed information patterns for both hedging and treatment terms, we remove a random 10% of terms that occurred at least once in the PCCRI corpus from each list and re-perform the analysis. This process is repeated with a different 10% of terms removed each time, until all terms have been removed at least two times.

In what is perhaps our most intriguing application of contextualization, we study

how the expression of distressing emotions by patients affects patterns of information flow, both locally (by comparing CODYM patterns between turns in which anger, fear or sadness were or weren't expressed) and globally (by comparing CODYM patterns between entire conversations in which anger and/or fear were or weren't expressed). As part of the latter analysis, a Random Forest classifier (implemented using Python's Scikit-learn package with 100 trees and default settings [46]) was trained to distinguish conversations that included anger and/or fear from those that didn't, using normalized conversation-level transition frequencies from $1^{st}$-order through $5^{th}$-order CODYMs as input features. Each Random Forest was trained on a random 80% of conversations from each class and tested on the remaining 20% of conversations from each class. This process was repeated 1,000 times for each order CODYM, with different random 80-20 data splits. The distribution of observed testing accuracies in these 1000 Random Forests was used to assess how consistently CODYM features could distinguish between conversations with fear and/or anger and those without.

### 4.2.4 The PCCRI corpus

The Palliative Care Communication Research Initiative (PCCRI) is a multisite observational cohort study conducted between January 2014 and May 2016 [43]. The study took place at two large U.S. academic medical centers, one in the Northeast and one in the West. Any English-speaking patients who were hospitalized and referred for inpatient palliative care consultation were eligible for this study, provided they were diagnosed with a metastatic nonhematologic cancer, did not have a documented exclusively comfort-oriented plan of care at the time of referral, were age 21 or over, and were able to consent for research either directly or via health care proxy (if lacking

capacity as determined by the clinical team). All members of the interprofessional Palliative Care Inpatient Consult teams at both sites were eligible to participate.

A total of 240 hospitalized patients with advanced cancer at the time of referral for inpatient palliative care consultation were enrolled in the study. Four withdrew, three died, and two were discharged before completing the palliative care consultation. Each consultation comprised one to three conversations between the patient, and potentially family members and/or close friends of the patient, and the palliative care team. More than one conversation occurred with the same patient when the initial conversation was only a preliminary assessment or when a conversation was interrupted prematurely (e.g., a patient was taken for x-rays) [43].

All conversations that were part of a palliative care consultation were audio recorded. With prior informed consent from all study participants, digital recorders were placed in unobtrusive locations in the rooms where the conversations took place (e.g., on a tray table next to a patient's bed); research assistants retrieved the recorders at the end of the visit by the palliative care team. All audio recordings were later transcribed verbatim and prepared in a standard format to facilitate natural language processing. The speaker during each transcribed turn was tagged as either a patient (which could include family members and/or close friends who were present in support of the patient) or a clinician, except in rare occasions ($< 1\%$ of turns) when transcribers could not determine whether the speaker was from the patient side or the clinician side (these turns were excluded from analyses that stratified turns by patient and clinician). In total, we examined 360 conversations that were recorded and transcribed for 231 unique patients.

Of these, five transcripts were excluded from this study because either a high

proportion of speaker turns were inaudible rendering the transcripts very incomplete, or because the conversations were too short (less than 20 speaker turns long) to perform meaningful analysis. The remaining 355 conversations, which were used in our analyses, contained 1,464,167 total words (14,812 unique) in 101,563 total speaker turns, with a median of 242 turns per conversation (Fig. 4.9).

Since transcribers were not always able to distinguish patients from family members or friends, or which clinician was speaking, in this work we adopt the convention that "patient" refers to anyone on the patient-side of the conversation and "clinician" refers to anyone on the clinical team. Although there were up to 8 participants in a conversation, patient turns were followed by clinician turns (and *vice versa*) during 86% of all transitions, closely resembling the alternating speaker pattern that would be expected in dyads.

All speaker turns in which the patient was audibly perceived to be expressing anger, fear, or sadness, had been previously labeled in the PCCRI transcripts, using well-established and reliable human coding methods [52, 53]. Anger was defined to include expressions of either frustration or anger. Fear was defined to be inclusive of words and sounds indicating worry, anxiety, fear or terror. Sadness was defined to include expressions with sad, disappointed, depressed, hopeless or discouraged sentiments. Turns that included multiple sentiments were coded as such. Ambiguous words or sounds that *might* indicate underlying emotion, or referred to emotions felt in the past, were not included in our analysis.

## 4.3 Results

### 4.3.1 Normative patterns of information flow in serious illness conversations

CODYMs reveal normative patterns ("fingerprints") of information flow in serious illness conversations. The existence of such normative patterns is supported by the uni-model distributions observed for all states and transitions in 3$^{rd}$-order CODYMs, for both patients and clinicians, indicating a prevailing pattern across all conversations in the corpus (Figs. 4.12, 4.13, and 4.14). These patterns are similar, and yet distinct, between patients and clinicians.

When interpreting the populated CODYMs, it is important to consider that only 42% of patient turns are long whereas 53% of clinician turns are long. Thus, to create appropriate null models, we randomize the locations of specific turns (thereby preserving the exact distributions of patient and clinician turn lengths), while maintaining the overall sequential order of patient and clinician turn-taking in the actual conversations. Null 3$^{rd}$-order CODYMs are generated from 1000 randomized versions of these conversations, stratified by patient and clinician turns.

Despite the skew apparent in most of these distributions (Figs. 4.12, 4.13, and 4.14), we elect to populate the CODYMs with the means (rather than medians) of transition and state values, since this preserves the more intuitive properties that (a) the sum of all edge weights in a CODYM is 100%, (b) the sum of all node weights in a CODYM is 100%, and (c) for each node, the sum of weights of incoming edges equals the sum of the weights of outgoing edges. We verified that using the means,

rather than the medians, does not qualitatively change any insights or conclusions drawn from the models (see Figs. 4.12, 4.13, and 4.14 for a comparison of mean *vs.* median values).

When comparing the observed CODYMs (Fig. 4.3, left column) to the expected values from the corresponding null models (Fig. 4.3, right column), states SLS and LSL, and the transitions between these two states, occur more frequently in the observed data than expected by chance in both patients and clinicians, whereas all other states occur less frequently than expected by chance (Tables 4.5 and 4.6). This indicates that conversations include more one-way information exchanges than would be expected by chance, and that sometimes it is the clinician imparting more information and sometimes it is the patient.

Despite the qualitative similarities between normative patterns of patients and clinicians described above, all but two of the eight states and five of the sixteen transitions are siginficantly different for the two speaker types (Tables 4.5 and 4.6). Most notably, it is evident that the LSL state is the most frequent state prior to a patient turn and is most often followed by a short patient turn (Fig. 4.3, upper left) whereas the SLS state is the most frequent state prior to a clinician turn and is most often followed by a long clinician turn (Fig. 4.3, lower left). This implies that clinicians, rather than patients, are most often imparting more information in one-way information exchanges. Other aspects of the normative CODYM patterns are discussed in subsequent sections.

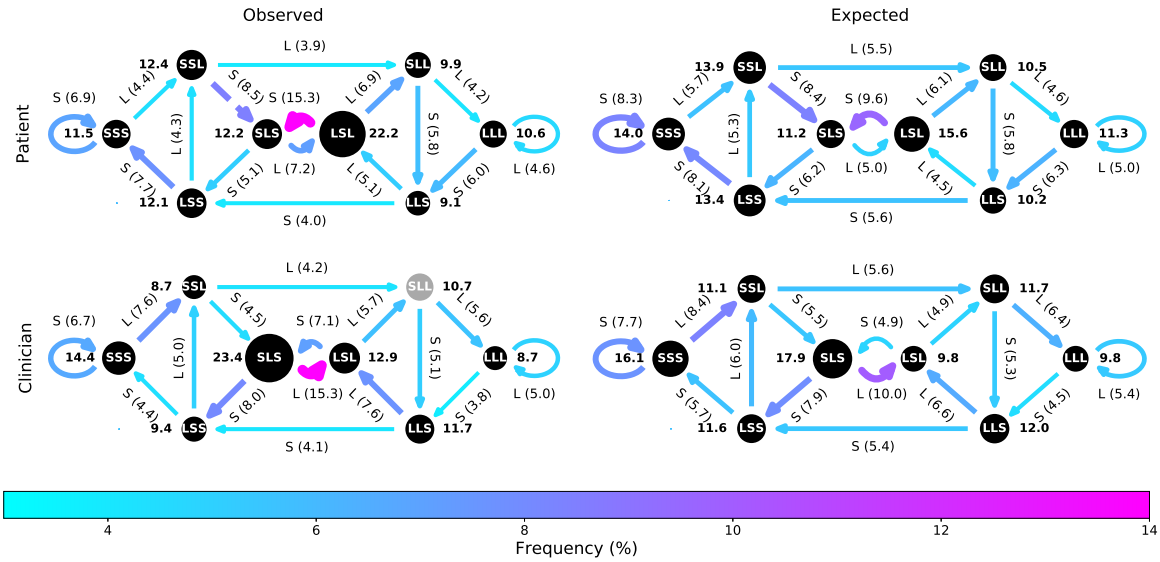



*Figure 4.3:* ***CODYMs of normative patterns in PCCRI corpus.*** *CODYMs of normative patterns of information flow for patient turns (top left) and clinician turns (bottom left), averaged over all conversations in the PCCRI corpus. Null models were constructed with the same turn length imbalance for patients and clinicians and the same sequential order of patient and clinician turns in each conversation (right column). Edge thickness and color indicate %Observed for each transition, as shown parenthetically on edge labels. Node diameter indicates %Observed for each state, as shown by the node labels in bold. All state and transition values were significantly different from their corresponding null models, according to the empirically derived 95% confidence intervals, with the exception of state* SLL *for clinicians (shown in gray).*

### 4.3.2 Dynamic Changes in Normative Patterns of Information Flow

Some transitions show distinct patterns of change in frequency across temporal deciles of conversations (Figs. 4.4, 4.15, and 4.16). For example, the short two-way information exchange, while occurring overall less frequently than expected by chance (Fig. 4.3), occurs more frequently in the first and last decile of the conversation for both patient turns (Fig. 4.4a) and clinician turns (Fig. 4.15). This is consistent with our observation

that words associated with ritualistic openings and closings of a conversation are over-represented in short two-way information exchanges (see Contextualization by Word Clustering, in the next Section). For patient turns, we also see a decrease in one-way information exchanges from patient-to-clinician from deciles 5 through 10 (Fig. 4.4b, $\rho = -0.95, p = 0.004$), but no corresponding significant increase in the one-way information flow from clinician-to-patient over deciles 5 through 10 (Fig. 4.4c, $p = 0.84$). The complementary changes in one-way information flow patterns for clinician turns exhibit nearly identical patterns (Figs. 4.15 and 4.16).

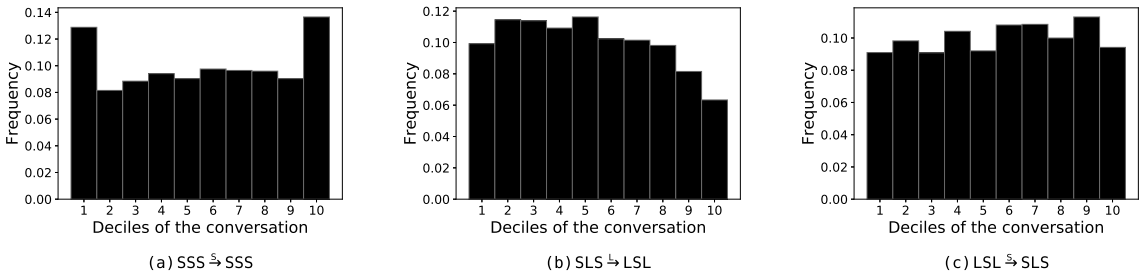



*Figure 4.4:* ***Selected temporal changes in CODYMs of PCCRI corpus.*** *Histograms of temporal changes in transition frequencies for patient turns in $3^{rd}$-order CODYMs for (a) short two-way information exchanges, (b) one-way information exchanges from patient-to-clinician, and (c) one-way information exchanges from clinician-to-patient, over 10 conversational deciles that were subsequently averaged over all conversations in the PCCRI corpus and normalized, such that the sum of all bins is 1.0. See Figs. 4.15 and 4.16 for plots of temporal changes in all transition frequencies, for both patient and clinician turns.*

### 4.3.3 Contextualization

#### Contextualization by word clustering

Unsupervised clustering was used to group 114 words (selected as previously described in the Methods) into six clusters, based on the transition frequencies of individual words in $2^{\text{nd}}$-order CODYMs. For each cluster, we generate a representative CODYM

by averaging the transition frequencies for all words in the cluster (for the distribution of transition frequencies in each cluster, see Fig. 4.17). We compare these to the CODYM of occurrences of all words (Fig. 4.2), to see which transitions are over- or under-represented. Because the CODYM of occurrences of all words is not formulated from a distribution of random trials, it is not possible to empirically derive the confidence interval to determine whether the differences between observed and expected frequencies are significant. However, the purpose of the clustering is not to determine which state/transition frequencies are significantly different from random, but to observe which words share similar patterns of information sharing.

We find that many words with similar information flow patterns fill similar functions in the context of information flow, and we refer to these clusters by monikers according to these functions (Table 4.1, Fig. 4.5). A complete listing of words in each cluster and their frequencies in the PCCRI corpus are shown in Table 4.7. Results are displayed in Fig. 4.5 as $\Delta Frequency = \%Observed - \%Expected$ to accentuate how each cluster differs from the norm. (The $\%Observed$ for each transition in each cluster are shown in Table 4.8.)

Three similar clusters include words that often occur in S turns during one-way information exchanges, linguistically referred to as "continuers" [54, 55, 56]. We refer to these three clusters as "strong", "moderate", and "weak" continuers, based on the degree to which the one-way information exchange pattern differs from that in the CODYM of all word frequencies. The most specialized grouping, the strong continuer cluster, comprises sounds of acknowledgment that serve to encourage the primary speaker to continue (e.g., hm, aha). These words have the most extreme specialization as continuers and are rarely used in other contexts, nearly always

*Table 4.1:* ***Word clusters from CODYMs of PCCRI corpus.*** *Six unsupervised clusters of 114 words, based on similarities in transition frequencies in a $2^{nd}$-order CODYM. For each cluster, we identify the number of words in the cluster, the percentage of turns that included any words in the cluster that were long (L), and example words in the cluster. See Table 4.7 for the complete lists of words in each cluster, and how many times each of these words appeared in the PCCRI corpus.*

| **Cluster Moniker** | **# Words** | **%L** | **Example Words** |
|---|---|---|---|
| Strong Continuers | 4 | 10.8 | hm, hmmm, mm, aha |
| Moderate Continuers | 10 | 36.9 | yeah, yes, yep, yup, ok, wow, huh |
| Weak Continuers | 19 | 63.5 | fine, right, sorry, no, beautiful, meeting, appreciate |
| Openers/Closers | 7 | 44.3 | thank(s), meet, hi/hello, welcome, bye |
| Clinical Talk | 51 | 95.5 | um, comfort, symptoms, treatments, chemotherapy, continue, disease |
| Potpourri | 23 | 84.8 | hurts, tylenol, concern, risk, confused, scary, funny |

occurring during S turns and, specifically, appearing in the $\mathsf{SL} \overset{s}{\Rightarrow} \mathsf{LS}$ transition more frequently than all other transitions combined. The moderate continuer group contains positive affirmations (e.g., yeah, yes, yep, ok) that, while often used as continuers, have a less extreme over-representation in the one-way information exchange, and not infrequently (in over a third of turns) occur in L turns. The weak continuer cluster contains some words that act as continuers in one-way information exchanges (e.g., right, no, nice, great, fine, exactly), but also occur in L turns nearly two thirds of the time and contain some words that aren't generally considered continuers (e.g., meeting, appreciate, appetite).

The "openings/closings" cluster contains words that are almost exclusively associated with starts and ends of conversations, occurring nine times as often in short two-way information exchanges, relative to the CODYM of all word frequencies. This is consistent with our earlier observation that short two-way information exchanges occur most often in the first and last deciles of conversations (Fig. 4.4).

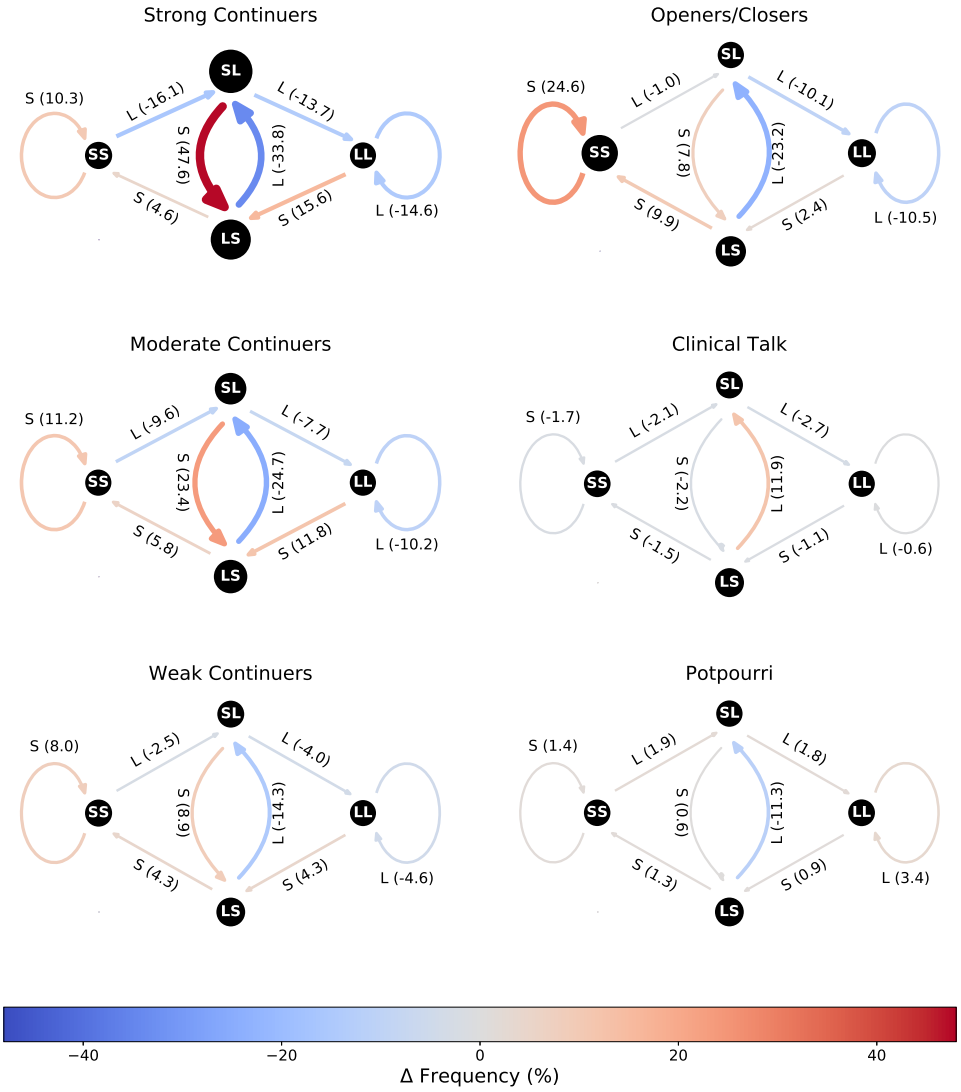



*Figure 4.5:* ***CODYMs of word clusters in PCCRI corpus.*** *Differences in $2^{nd}$-order CODYM information flow patterns of six clusters of words created by unsupervised clustering (Table 4.1), relative to expected frequencies based on all words in the PCCRI corpus. Edge thickness, color, and labels indicate the amount by which the frequencies differ from those of all words in the corpus, $\Delta Frequency = \%Observed - \%Expected$, where $\%Expected$ is as shown in Fig. 4.2. Node diameter is proportional to $\%Observed$.*

The cluster we have dubbed "clinical talk" is the largest cluster, containing 51 of the 114 words considered. The words in this cluster most often occur during L

turns in one-way information exchanges, with a slight under-representation in all other transitions, relative to the model of all word frequencies. While many of the words in this cluster specifically relate to clinical talk (e.g., comfort, symptoms, treatments, chemotherapy), others are more general (e.g., sort, may, whether); the most prevalent word in this cluster is "um", a word often used for holding the floor during a L turn.

The remaining 23 words were grouped into a cluster we refer to as "potpourri", since the relationships between these words are less obvious than in the other clusters. Although several of them described feelings (e.g., hurts, concern, confused, scary, funny), the word risk, a word used predominantly by clinicians (Table 4.3), also appears in this cluster. The general pattern of information flow in this cluster is opposite of what was observed for the "clinical talk" cluster. Specifically, these words are under-represented on the L turns of one-way information exchanges and slightly over-represented on all the other transitions, more so on other L turns than on S turns, with the greatest over-representation during long two-way exchanges.

### Contextualization by hedging and treatment terms

To assess whether patients and clinicians use hedging and/or treatment terms differently, we stratify the corpus into patient turns and clinician turns. We compare observed CODYM transition frequencies to expected frequencies in null CODYMs, where these null models were created from size-matched samples drawn according to the known frequencies of each state/transition over all patient and clinician turns. In this way, the null models capture the amount of random noise that would be expected due to chance, given the sample size. As shown in Fig. 4.20, the mean frequencies of these stratified null CODYMs are very similar to the single CODYM of the entire

unstratified corpus already shown in Fig. 4.2, so are not shown in the main text.

Patients and clinicians used hedging terms 3,805 and 5,794 times, respectively. Overall, hedging terms increased in frequency over the first half of the conversation, especially when used by clinicians (for patients, the usage of hedging terms peaked in decile 3); the use of hedging terms then remained high until decile 9 for both speaker types, dropping in the last decile (Fig. 4.11). The differences in CODYM patterns where hedging terms were used, relative to the CODYMs of all words, were similar for patients and clinicians (Figure 4.6, left column). Hedging terms are generally used 2-3% more frequently than expected in L turns following another L turn, especially during long two-way information exchanges. Patients, but not clinicians, used hedging terms over 3% less frequently in the L turn of a one-way information exchange, than in the corresponding null models. All other transitions with significant differences (Fig. 4.6, left column, solid transition arrows) had hedging terms slightly under-represented, relative to the corresponding null models.

Patients and clinicians used treatment terms 4,247 and 10,469 times, respectively. In contrast to the relatively similar CODYM patterns in their use of hedging terms, patients and clinicians exhibited very different CODYM patterns in their use of treatment terms (Figure 4.6, right column). Clinicians used treatment terms nearly 6% more frequently during L turns in one-way information exchanges (from clinician-to-patient), with $\leq 1.3\%$ absolute differences from expected for all other transitions, nearly all significantly lower than expected (Figure 4.6, lower right, solid transition arrows). Patients, however, used treatment terms over 3% more frequently than expected on the SL$\overset{L}{\rightarrow}$LL transition, but with $\leq 1.3\%$ absolute differences from expected for all other transitions, most of which were not significantly different from the null

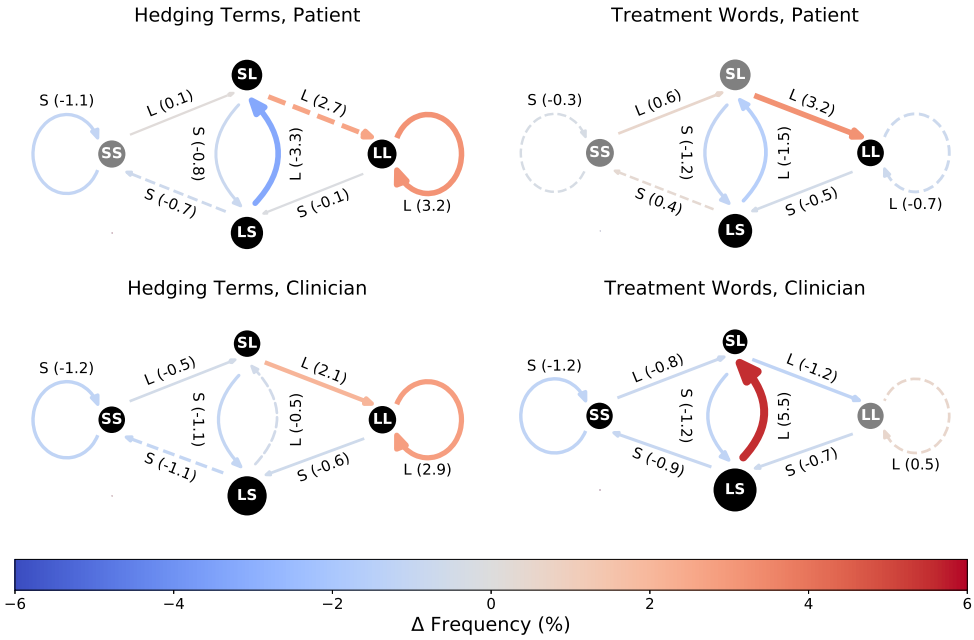



*Figure 4.6:* ***CODYMs for hedging and treatment terms in PCCRI corpus.*** *Differences in $2^{nd}$-order CODYM patterns for hedging and treatment terms, relative to the means of null models sampled from all words in the corpus, stratified by patient and clinician turns. Edge labels indicate the length of the turn in the transition (S vs. L), parenthetically followed by the observed discrepancy in the percentage occurrence. Dashed transition arrows, and nodes colored gray, represent transitions and states where observed frequencies were not significantly different from expected, according to the empirically derived 95% confidence intervals. Edge thickness, colors, and labels represent $\Delta Frequency = \%Observed - \%Expected$. Node diameter is proportional to %Observed.*

model (Figure 4.6, upper right, dashed transition arrows).

The overall CODYM patterns shown in Fig. 4.6 were not qualitatively affected by removal of different random sets of 10% of hedging terms or treatment terms. This robustness to removal of random terms confirms that the observed patterns within term groupings are not reliant on a particular subset of terms.

**Contextualization by expression of distressing emotion**

In total, 208 conversations (58.6%) contained at least one patient turn where anger, fear or sadness was detected, and in 102 conversations (28.7%) at least two of these different types of distressing emotion were present. Conversations that had at least one patient turn with a distressing emotion typically (89.4% of the time) had more than one, with a heavy-tailed distribution indicating that a small number of conversations had very high instances of emotion. It was also common to have multiple consecutive patient turns where emotion was expressed. Consequently, the previous turns, used in defining the state prior to a given transition where emotion was present, also often included expressed emotions. This should be considered when interpreting the results herein.

The expression of anger, fear, and sadness occurs during L patient turns 69.8%, 66.6%, 56.7% of the time, respectively, whereas only 42% of all patient turns were L (Fig. 4.18). To compare CODYM patterns in patient turns with or without expression of distressing emotions, we analyzed all patient turns with audibly expressed anger, fear or sadness, relative to null CODYMs created by randomly sampling patient turns in which no distressing emotion was present, but with the same number of patient turns and same frequency of L turns as in those patient turns where distressing emotion was detected. As a result of the higher proportion of patient L turns, the resulting null CODYM models for expression of distressing emotion differ from the normative CODYM pattern of discourse for all patient turns (compare the models in Fig. 4.7, right column with Fig. 4.3, top right).

The observed CODYM patterns during turns where distressing emotion was expressed (Fig. 4.7, left column) differ markedly from what would be expected due to

the increased proportion of patient long turns alone (Fig. 4.7, right column). Note that, for anger and fear, all but one state and most transitions differ significantly from their respective null model. However, for sadness, only three states and fewer than half of transitions are significantly different from the null model. The most notable differences, present in all three emotions, are (a) an over-representation of expression of distressing emotions during L patient turns in one-way patient-to-clinician information exchanges, and (b) an under-representation during S patient turns in one-way clinician-to-patient information exchanges. These differences are most extreme for turns in which anger is expressed, and least extreme when sadness is expressed. Additionally, fear is significantly under-represented during short two-way information exchanges, relative to its null model. Expressions of anger, fear, and sadness all peak in deciles 4 and 5 (Fig. 4.19), and then decrease over the remainder of narrative time in the conversations. The strong similarities in temporal patterns of the three emotions arise, in part, because multiple turns were determined to include more than one of these emotions.

To examine the impact of distressing emotions on information flow at the conversation level, we focused on expressions of anger and fear, since expressions of sadness caused relatively few significant differences at the turn level (as seen in Fig. 4.7). Out of 355 conversations, 187 (53%) had at least one instance of anger or fear, creating a relatively balanced data set. The distributions of CODYM transition frequencies for conversations with at least one instance of anger or fear differed significantly from the distributions of CODYM transition frequencies for conversations without any expression of fear or anger, for 13 of 16 transitions (2-sample Kolmogorov-Smirnov test, $p < 0.05$, Fig. 4.8a solid edges). In conversations with at least one expression of anger

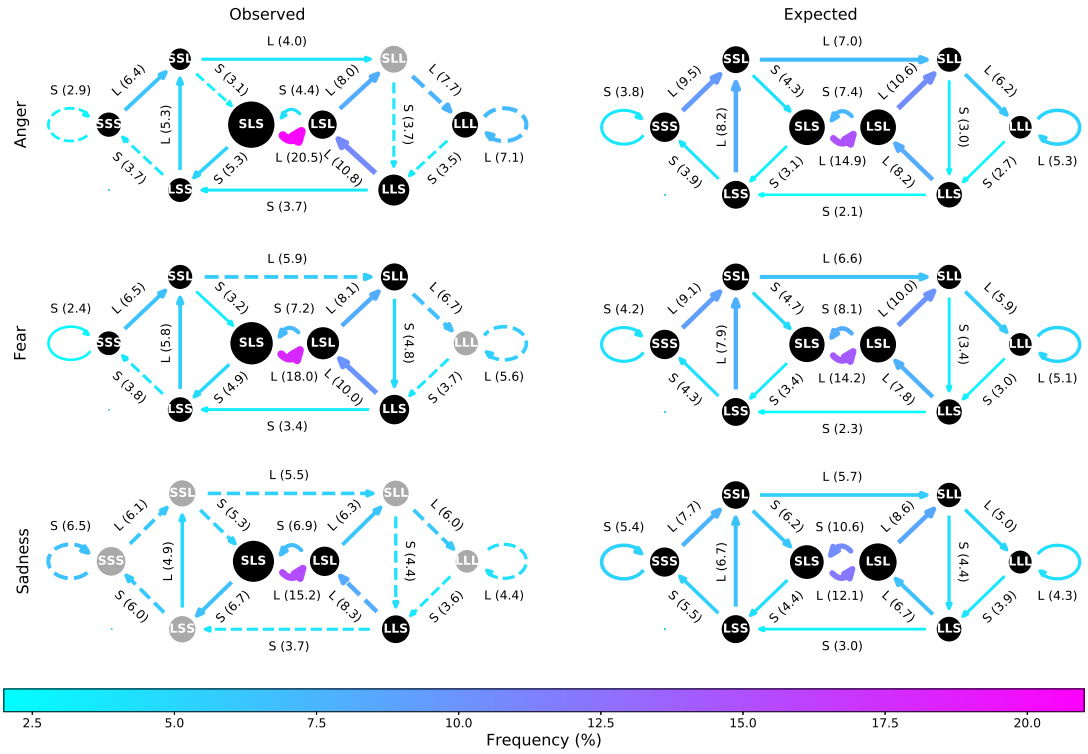
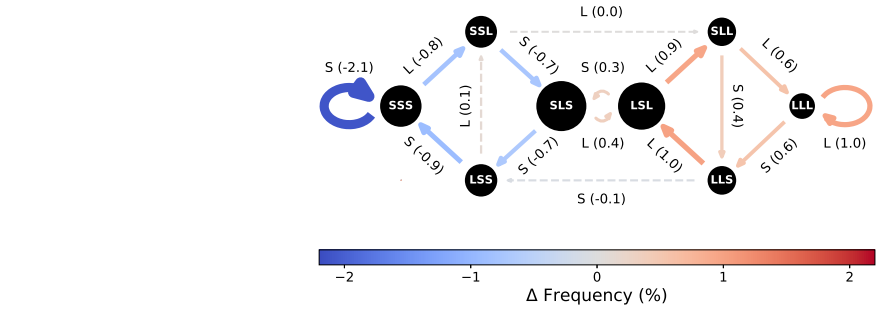



*Figure 4.7:* ***Turn-level CODYMs by emotional content in PCCRI corpus.*** *Observed and null CODYMs for patient turns with audibly perceptible expressions of distressing emotion (anger, fear, and sadness) for the PCCRI corpus. Edge thickness and color indicate %Observed for each transition, as shown parenthetically on edge labels. Node diameter indicates %Observed for each state. Dashed transition arrows, and nodes colored gray, represent transitions and states where observed frequencies were not significantly from expected, according to the empirically derived 95% confidence intervals.*

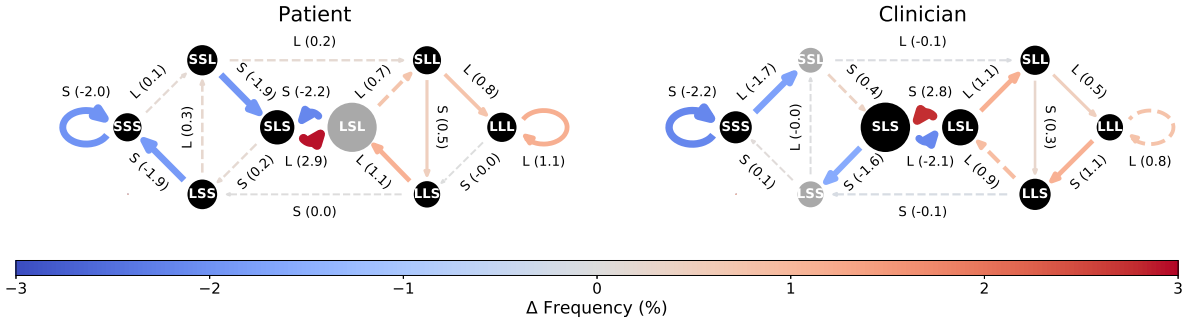
or fear, transitions through states with a predominance of L turns, including long two-way information exchanges, and transitions that perpetuate one-way information exchanges, are over-expressed (shown in warm colors in Fig. 4.8a). Conversely, transitions through states dominated by S, especially short two-way information exchanges, are under-expressed (shown in cool colors in Fig. 4.8a). CODYMs from conversations that are stratified by patient and clinician turns tell a more nuanced story. Not only do these show a larger range in absolute differences of transition frequencies between

conversations with and without anger or fear than in the unstratified CODYM model (note the different ranges on the colorbars in Figs. 4.8a,b), but there are complementary changes in patients and clinicians. Most notably, conversations where anger or fear are expressed have more one-way information flow from patients to clinicians (Fig. 4.8b, left) and less one-way information flow from clinicians to patients (Fig. 4.8b, right), than do conversations with no anger or fear expressed. In addition, we see that the patient extends long two-way information exchanges significantly more often in conversations where they are expressing anger or fear.

The differences in CODYM transition frequencies between conversations with or without expressions of anger or fear were significant enough to produce a signal with predictive power. Random Forest classifiers, trained on conversation-level transition frequencies of $1^{\text{st}}$- through $5^{\text{th}}$-order CODYMs, were able to predict which conversations included any expression of anger and/or fear more often than expected by chance ($p < 0.05$, Table 4.2). Prediction accuracy (averaged over 1000 trained classifiers) was (a) slightly higher when using data from stratified *vs.* unstratified CODYMs, and (b) generally increased with increasing CODYM order. For example, the $3^{\text{rd}}$-order stratified transitions frequencies (shown in Fig. 4.8) yielded an average prediction accuracy of 64.5%, whereas $5^{\text{th}}$-order stratified transition frequencies raised that to 67.3%. Even $1^{\text{st}}$-order stratified transition frequencies exhibited an average prediction accuracy of nearly 60%.




(a) Unstratified Models




(b) Stratified Models

*Figure 4.8:* ***Conversation-level CODYMs by emotional content in PCCRI corpus.*** *Differences between conversations with and without at least one expression of anger or fear, for 3rd-order CODYMs where (a) the data are not stratified by speaker role, and (b) each conversation is stratified by patient and clinician turns. Edge labels indicate the length of the turn in the transition (S vs. L), parenthetically followed by the observed difference in the percentage occurrence. Dashed transition arrows, and nodes colored gray, represent transitions and states where observed frequencies were not significantly different from expected, according to the empirically derived 95% confidence intervals. Edge thickness, colors, and labels represent* $\Delta Frequency = (\%Observed\ with\ anger\ or\ fear) - (\%Observed\ without\ anger\ or\ fear)$. *Node diameter is proportional to %Observed with anger or fear present.*

## 4.4 Discussion

We have shown how CODYM analysis enables one to quantify, visualize, and compare high-level patterns in conversational dynamics from one or more conversations. We have made a deliberate choice to keep CODYMs simple, requiring only sequences of binarized

turn lengths and using memory of only 2-3 turns. We have made these choices in order to facilitate interpretability, ensure adequate sample sizes, and to protect privacy. Because of this, there are many important conversational features that CODYMs, as defined here, do not directly incorporate, such as interruptions, overlapping speech, conversational pauses, accent, intonation, gestures, facial expressions and eye contact, all of which play some role in conversational discourse (e.g., see [57] for a discussion of the importance of "connectional silences" in serious illness conversations). Although any of the aforementioned features *could* be directly incorporated into a CODYM, the state space of the resulting model would be larger, less well populated, and more difficult to interpret. As illustrated by our analyses of hedging terms, treatment terms, and expression of distressing emotion, even simple turn-length-based CODYMs provide a means to consider additional verbal and non-verbal features, through contextualization.

CODYM visualizations effectively summarize the overall patterns of sequential dependencies in speaker turn lengths in concise plots, whose size and complexity are independent of the number of turns being analyzed. This facilitates rapid identification and comparison of patterns of information flow in sequences of turns that may include all or parts of one or many conversations. Since CODYM construction does not require access to detailed conversational content, the approach completely preserves the privacy of conversational participants and can conceptually be incrementally constructed in real-time, precluding the need for storage, or even transcription, of conversations (as discussed further in Conclusions and Future Work). Although CODYMs can be contextualized and can be used to study temporally discretized patterns in information flow, they are not natively designed to capture the turn-level

flow of concepts in individual conversations. CODYMs and Discursis [29] are thus complementary methods for visualizing information flow in conversations at different scales and levels of detail.

Through unsupervised word clustering, we discovered that related groups of words often exhibit similar CODYM patterns. This, in itself, helps to validate that CODYMs represent semantically meaningful patterns of information flow. Three of the emergent clusters contained words that are specialized, to different degrees, for use in the short turns of one-way information exchanges, consisting primarily of strong, moderate, and weak continuers. This is consistent with conversation analytic theory, which suggests continuers are words or phrases used by speakers to convey they are relinquishing an opportunity to command the speaking floor, thus signaling the currently dominating speaker to continue [54, 55, 56]. Words in the strong and moderate continuer clusters alone appear 44,548 times in the PCCRI corpus (Table 4.7), accounting for 41.8% of all short turns in the corpus. The short two-way information exchange appears less frequently than expected by chance (Fig. 4.3), with one notable exception. Word clustering identified a group of openers and closers that occur disproportionately often during short two-way information exchanges (Fig. 4.5). The importance of openers and closers in conversation was explored by some of the early pioneers of conversation analysis [8, 58] and has been the subject of much attention since then [59, 60, 61, 62, 63]. These ritualistic transitions into and out of a conversation serve an important role, although little information need actually be exchanged. It is thus to be expected that openers and closers appear most often in short two-way exchanges (Fig. 4.5), and that this information flow pattern occurs most commonly at the beginnings and ends of conversations (Fig. 4.4a).

Conversation has generally been observed to be asymmetric, with a single dominant speaker holding the floor much of the time [64]. During these one-way information exchanges in didactic conversations, the typical alternating pattern of speakers results in one speaker taking all the long turns (being the "talker") while the other takes the short turns (being the "listener"). This is particularly true in institutional settings with defined speaker roles [64], so we expect it will be prevalent in clinical conversations, given the power dynamics between clinicians and patients. Consistent with this expectation, one-way information exchanges from clinician-to-patient were observed to be the most common information sharing pattern in the PCCRI corpus (Fig. 4.3), and occurred throughout serious illness conversations at a roughly constant level (Fig. 4.4c). It is likely that these one-way clinician-to-patient information exchanges are often used for delivering information and treatment options related to the patient's illness. Thus, it is not surprising that treatment terms are most often used when clinicians are in the talker role (Fig. 4.6), and words commonly used in clinical talk clustered together as words occurring most often during one-way information exchanges (Fig. 4.5), presumably when the clinician is the talker.

Overall, patients adopt the talker role less often than clinicians do, although still more frequently than expected by chance, in serious illness conversations (Fig. 4.3). In pooled patient turns, over all conversations in the corpus, one-way information exchanges from patient-to-clinician are most frequent near the centers of conversations and then steadily decline (Fig. 4.4b). These temporal changes in patient as talker are similar to the temporal patterns in the usage of treatment terms reported in [45], and also roughly coincide with temporal patterns in the expression of distressing emotions (Fig. 4.19). We also note that patient expressions of distressing emotion

occur disproportionately often in long turns (Fig. 4.18), in particular during one-way patient-to-clinician information exchanges (Fig. 4.7). In our previous work, we observed that sentiment scores become increasingly positive over the course of the PCCRI conversations, which was largely attributed to a decrease in the use of disease-related terms that are ascribed negative sentiments [45]. We hypothesize that the palliative care clinicians in the PCCRI, trained to be highly skilled communicators, may be taking on the role of "good listener" [64] during portions of the conversation in which difficult topics are being discussed, as a means of encouraging the patients to express their values and preferences relating to available treatment options. After the patient has finished expressing themselves and one-way patient-to-clinician information exchanges abate, these serious illness conversations may then naturally come to a close.

We also observed an increase in the overall frequency of one-way patient-to-clinician information exchanges, along with a decrease in the prevalence of short two-way information exchanges and an increase in long two-way information exchanges, in conversations where fear or anger are expressed (Fig. 4.8). While these differences at the conversation level are modest, they are strong enough to have predictive power in distinguishing conversations with anger or fear from those without. This is especially true when the data are stratified by patient and clinician turns; but even when the speaker role is completely anonymized we find that differences in information flow patterns can be a marker for expression of distressing emotions. Our findings regarding the impact of distressing emotions on patterns of conversational dynamics are important because anger, fear, and sadness manifest frequently in palliative care conversations [53], and have been found to have therapeutic effects. Being able to

process and express distressing emotions is linked to improved health outcomes for patients with cancer [65], and it has recently been shown in the PCCRI corpus that expression of anger, in particular, is associated with improvement in how much patients feel heard and understood by the clinical team following a palliative care consult [44].

Uncertainty is prevalent in the discussion of prognosis [66] and treatment options [67], where outcomes are often unknown. Effectively communicating this uncertainty is essential to having patients fully informed before they make decisions regarding their care [68, 69]. Helping patients to manage uncertainty is fundamental to patient-centered communication and may be a key ingredient in the beneficial effects of palliative care consultations on patient quality of life [67, 70, 71]. Hedging terms are often used to soften claims and insert uncertainty into a statement, and are the most prevalent subtype of uncertainty found in the PCCRI corpus [51]. As we found previously, the general story arc of conversations in the PCCRI corpus moves from discussion of symptoms, to treatments, to prognosis, to use of modal verbs, peaking in deciles 2,4,6,9, respectively [45]. Here, we observed that the use of hedging terms by clinicians is highest from deciles 5-9, coinciding with the parts of the conversations where prognosis terms and modal verbs (indicators of possibility) are most frequent. This suggests that clinicians may be using hedging terms in conveying prognostic uncertainty.

Long two-way information exchanges represent conversational regimes where alternating speakers are trading information. Although this is one of the least frequent information sharing patterns in the corpus (Fig. 4.3), it is more frequent in conversations where audible patient expressions of anger and fear occur (Fig. 4.8) and in turns

when commonly used words associated with feelings are expressed (Fig. 4.5). It is thus interesting that long two-way exchanges are also over-represented when hedging terms are used, for both patients and clinicians (Fig. 4.6). This may indicate that there is bilateral sharing regarding uncertainty about difficult topics, as would be expected to occur when clinicians are actively helping patients to manage their uncertainty.

## 4.5 Conclusions and future work

We have presented and validated a novel approach to quantify and visualize the dynamics of information flow in conversations with CODYMs (COnversational DYnamics Models). CODYMs are the first Markov Model to use speaker turn length as the fundamental unit of information and to provide concise, high-level, quantitative summaries of overall dependencies in sequences of speaker turn lengths. This new approach facilitates identification and comparison of normative patterns of information flow across sequences of turns from one or more conversations, in context-independent or context-dependent ways. CODYMs complement existing qualitative and quantitative approaches for studying conversational dynamics, and comprise a new tool for conversational analysis. We provide open source code for populating, visualizing, and contextualizing CODYMs [72].

We applied the method to a unique and important corpus of palliative care consultations with seriously ill patients. We discovered normative patterns of information flow in these conversations that differ between patients and clinicians, and between conversations with and without expressions of distressing emotions. While these normative patterns are interesting in their own right, they may also have practical

applications. For example, it would be interesting to compare normative CODYMs from in-person palliative care consultations to those conducted remotely, to see how telehealth platforms impact the patterns of information flow between patients and clinicians. Similarly, CODYM analysis of mock consultations between clinicians-in-training and actors portraying seriously ill patients could be used to assess how closely the information flow patterns of these training scenarios reflect those observed in palliative care consultations with real patients. If associations can be found between CODYM patterns and quality indicators of healthcare conversations (e.g., the degree to which patients feel heard and understood [73], a measure that is currently being considered for widespread use [74]), these could provide valuable insights for institutions seeking to improve the quality of conversations with seriously ill patients.

Achieving patient-centered care, in which clinical decisions are guided by patient preferences, needs, and perspectives [75], has long been recognized as a means for improving healthcare delivery [76]. Patient-centered communication, which includes recognizing and responding to emotions, helping patients to manage uncertainty, promoting reciprocal exchanges of information to create a shared understanding, and helping patients make informed decisions regarding their care, is central to achieving this aim [77]. However, it is not clear how to effectively measure or assess patient centered communication [67]. In this work, we have shown how CODYM analysis offers a new approach to quantifying and visualizing patterns of information flow related to markers of patient centered communication, such as the rates of one-way patient-to-clinician information exchanges and long two-way information exchanges, and how these rates change when distressing emotions or uncertainty are expressed. CODYMs may thus be helpful in assessing the degree to which clinical conversations

are patient centered. For example, such an analysis could potentially be useful in understanding observed racial and ethnic disparities in prognosis communication [78].

It is not clear whether the conversational "fingerprints" uncovered in the PCCRI corpus are unique to serious illness conversations, or represent more general conversational paradigms in healthcare or other contexts. We suspect that the frequency of one-way patient-to-clinician information exchanges, and long two-way information exchanges, may be higher in serious illness communication relative to conversations in other clinical contexts (e.g., [79]). It will be fascinating to compare CODYMs across a wide variety of corpora from different languages, cultures, and contexts (including online conversations), to reveal which patterns of information flow in conversations are universal, and which are unique to certain settings.

Conversation analysis has traditionally been a discipline reliant on manual transcription of conversations with highly detailed annotations [6]. This is a resource-intensive process that requires full access to the often very private content of conversations. Indeed, the CODYM analyses presented here used transcriptions of audio-recordings of sensitive serious illness conversations. However, we envision an alternate formulation of CODYMs that uses turn duration (in seconds), in lieu of the number of words, for defining turn lengths. A time-based definition of turn length would facilitate real-time automation and analysis of conversational dynamics, precluding the need for transcription or even storage of conversational audio, thus completely protecting privacy. Large numbers of conversations are already taking place in a medium that is natively capable of capturing conversational data appropriate for automated CODYM analysis. For example, many popular video conferencing services already incorporate tools that automate the detection of speaker turns, and such services have exploded

in popularity in the wake of the Covid-19 global pandemic. Ongoing advances in the automated detection of conversational features including speaker recognition [80, 81], emotion [82, 83], conversational pauses [57], empathy [84, 85], gaze patterns [86], and word recognition [87], will facilitate real-time analysis and contextualization of CODYMs. Ultimately, we foresee a fully-automated pipeline for CODYM analyses, with no compromise to the privacy of conversational content.

As more conversational data become available, whether as transcriptions or through real-time processing, CODYMs will be a valuable tool for studying information flow in a wide variety of contexts and contributing to our understanding of how to have more effective conversations. Such a tool could be of practical utility in training and assessment of high quality communication in healthcare and other application domains, while also yielding new theoretical insights into conversational dynamics across languages, cultures, and contexts.

# Bibliography

[1] Carlo Prevignano and Paul J Thibault. *Discussing conversation analysis: The work of Emanuel A. Schegloff*. John Benjamins Publishing, 2003.

[2] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier, 1978.

[3] Enrico Coiera. When conversation is better than computation. *Journal of the American Medical Informatics Association*, 7(3):277–286, 2000.

[4] Richard L Street Jr, Gregory Makoul, Neeraj K Arora, and Ronald M Epstein. How does communication heal? pathways linking clinician–patient communication to health outcomes. *Patient Education and Counseling*, 74(3):295–301, 2009.

[5] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[6] Ian Hutchby and Robin Wooffitt. *Conversation analysis*. Polity, 2008.

[7] Elliott M Hoey and Kobin H Kendrick. Conversation analysis. *Research methods in psycholinguistics: A practical guide*, pages 151–173, 2017.

[8] Emanuel A Schegloff. Sequencing in conversational openings 1. *American Anthropologist*, 70(6):1075–1095, 1968.

[9] Galina B Bolden. Transcribing as research:"manual" transcription and conversation analysis. *Research on Language and Social Interaction*, 48(3):276–280, 2015.

[10] Michael A Westerman. Conversation analysis and interpretive quantitative research on psychotherapy process and problematic interpersonal behavior. *Theory & Psychology*, 21(2):155–178, 2011.

[11] Gabriele Kasper and Johannes Wagner. Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34:171–212, 2014.

[12] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[13] Marco Cristani, Anna Pesarin, Carlo Drioli, Alessandro Tavano, Alessandro Perina, and Vittorio Murino. Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8):1785–1800, 2011.

[14] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, pages 191–198. ACM, 2005.

[15] HB Kekre, CL Saxena, and HM Srivastava. A two-state markov model of speech in conversation and its application to computer communication systems. *Computers & Electrical Engineering*, 4(2):133–141, 1977.

[16] Kevin C Parker. Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965, 1988.

[17] Andrew P Thomas, Peter Bull, and Derek Roger. Conversational exchange analysis. *Journal of Language and Social Psychology*, 1(2):141–155, 1982.

[18] Mark G Core and James Allen. Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, volume 56, pages 28–35. Boston, MA, 1997.

[19] Andrew P Thomas. Conversational routines: A markov chain analysis. *Language & Communication*, 5(4):287–296, 1985.

[20] Andrew Thomas. Describing the structure of conversation using markov chains, chain analysis and monte carlo simulation. *Journal of Literary Semantics*, 16(3):159–181, 1987.

[21] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[22] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, Alex Pentland, et al. Learning human interactions with the influence model. NIPS, 2001.

[23] Alex Pentland. Social dynamics: The voice of power and influence. *Tech. Rep. 579*, 2004.

[24] A Pentland. Learning communities—understanding information flow in human networks. *BT Technology Journal*, 22(4):62–70, 2004.

[25] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 271–278, 2007.

[26] Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. A taxonomy and survey of dynamic graph visualization. In *Computer Graphics Forum*, volume 36, pages 133–159. Wiley Online Library, 2017.

[27] Andreas Wulvik, Matilde Bisballe Jensen, and Martin Steinert. Temporal static visualisation of transcripts for pre-analysis of video material: Identifying modes

of information sharing. *Anal. Des. Think. Stud. Cross-Cult. Co-Creat. Leiden CRC Press. Francis*, 2017.

[28] Tony Bergstrom and Karrie Karahalios. Seeing more: visualizing audio cues. In *IFIP Conference on Human-Computer Interaction*, pages 29–42. Springer, 2007.

[29] Daniel Angus, Andrew Smith, and Janet Wiles. Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):988–997, 2011.

[30] James J Thomas and Kristin A Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[31] Riccardo Fusaroli and Kristian Tylén. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40(1):145–171, 2016.

[32] Hai Ming Wong, Susan Margaret Bridges, Kuen Wai Ma, Cynthia Kar Yung Yiu, Colman Patrick McGrath, and Olga A Zayts. Advanced informatics understanding of clinician-patient communication: A mixed-method approach to oral health literacy talk in interpreter-mediated pediatric dentistry. *PloS One*, 15(3):e0230575, 2020.

[33] Daniel Angus, Bernadette Watson, Andrew Smith, Cindy Gallois, and Janet Wiles. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PloS One*, 7(6):e38014, 2012.

[34] Daniel Angus, Sean Rintel, and Janet Wiles. Making sense of big text: a visual-

first approach for analysing text data using leximancer and discursis. *International Journal of Social Research Methodology*, 16(3):261–267, 2013.

[35] Bernadette M Watson, Daniel Angus, Lyndsey Gore, and Jillann Farmer. Communication in open disclosure conversations about adverse events in hospitals. *Language & Communication*, 41:57–70, 2015.

[36] Rosemary Baker, Daniel Angus, Erin R Smith-Conway, Katharine S Baker, Cindy Gallois, Andrew Smith, Janet Wiles, and Helen J Chenery. Visualising conversations between care home staff and residents with dementia. *Ageing and Society*, 35(2):270, 2015.

[37] Paul Haidet and Debora A Paterniti. Building a history rather than taking one: A perspective on information sharing during the medical interview. *Archives of Internal Medicine*, 163(10):1134–1140, 2003.

[38] Rachelle E Bernacki and Susan D Block. Communication about serious illness care goals: a review and synthesis of best practices. *JAMA Internal Medicine*, 174(12):1994–2003, 2014.

[39] James A Tulsky, Mary Catherine Beach, Phyllis N Butow, Susan E Hickman, Jennifer W Mack, R Sean Morrison, Richard L Street, Rebecca L Sudore, Douglas B White, and Kathryn I Pollak. A research agenda for communication between health care professionals and patients living with serious illness. *JAMA internal medicine*, 177(9):1361–1366, 2017.

[40] Institute of Medicine (US). Committee on Approaching Death: Addressing Key

End-of Life Issues. *Dying in America: Improving quality and honoring individual preferences near the end of life*. National Academies Press, 2015.

[41] Betty Ferrell, Stephen R Connor, Anne Cordes, Constance M Dahlin, Perry G Fine, Nancy Hutton, Mark Leenay, Judy Lentz, Judi Lund Person, Diane E Meier, et al. The national agenda for quality palliative care: the national consensus project and the national quality forum. *Journal of Pain and Symptom Management*, 33(6):737–744, 2007.

[42] Padhraig Ryan, Saturnino Luz, Pierre Albert, Carl Vogel, Charles Normand, and Glyn Elwyn. Using artificial intelligence to assess clinicians' communication skills. *BMJ*, 364:l161, 2019.

[43] Robert Gramling, Elizabeth Gajary-Coots, Susan Stanek, Nathalie Dougoud, Heather Pyke, Marie Thomas, Jenica Cimino, Mechelle Sanders, Stewart C Alexander, Ronald Epstein, et al. Design of, and enrollment in, the palliative care communication research initiative: a direct-observation cohort study. *BMC Palliative Care*, 14(1):40, 2015.

[44] R Gramling, J Straton, LT Ingersoll, LA Clarfeld, L Hirsch, CJ Gramling, BN Durieux, DM Rizzo, MJ Eppstein, and SC Alexander. Epidemiology of fear, sadness and anger expression in palliative care conversations. *Journal of Pain and Symptom Management*, 2020.

[45] Lindsay Ross, Christopher M Danforth, Margaret J Eppstein, Laurence A Clarfeld, Brigitte N Durieux, Cailin J Gramling, Laura Hirsch, Donna M Rizzo, and Robert Gramling. Story arcs in serious illness: Natural language processing features of

palliative care conversations. *Patient Education and Counseling*, 103(4):826–832, 2020.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Veronika Vincze. *Uncertainty detection in natural language texts*. PhD thesis, szte, 2015.

[48] Yulia A Strekalova and Vaughan S James. Language of uncertainty: The expression of decisional conflict related to skin cancer prevention recommendations. *Journal of Cancer Education*, 32(3):532–536, 2017.

[49] Paul KJ Han, William MP Klein, and Neeraj K Arora. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838, 2011.

[50] David A Hanauer, Yang Liu, Qiaozhu Mei, Frank J Manion, Ulysses J Balis, and Kai Zheng. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *AMIA Annual Symposium Proceedings*, volume 2012, page 321. American Medical Informatics Association, 2012.

[51] Brigitte N Durieux. Uncertainty corpus, Oct 2019.

[52] Luke T Ingersoll, Stewart C Alexander, Susan Ladwig, Wendy Anderson, Sally A Norton, and Robert Gramling. The contagion of optimism: The relationship

between patient optimism and palliative care clinician overestimation of survival among hospitalized patients with advanced cancer. *Psycho-oncology*, 28(6):1286–1292, 2019.

[53] Stewart C Alexander, Susan Ladwig, Sally A Norton, David Gramling, J Kelly Davis, Maureen Metzger, Jane DeLuca, and Robert Gramling. Emotional distress and compassionate responses in palliative care decision-making consultations. *Journal of Palliative Medicine*, 17(5):579–584, 2014.

[54] Gail Jefferson. Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm". 1984.

[55] Emanuel A Schegloff. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing Discourse: Text and Talk*, 71:93, 1982.

[56] Victor H Yngve. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578, 1970.

[57] Brigitte N Durieux, Cailin J Gramling, Viktoria Manukyan, Margaret J Eppstein, Donna M Rizzo, Lindsay M Ross, Aidan G Ryan, Michelle A Niland, Laurence A Clarfeld, Stewart C Alexander, et al. Identifying connectional silence in palliative care consultations: a tandem machine-learning and human coding method. *Journal of Palliative Medicine*, 21(12):1755–1760, 2018.

[58] Emanuel A Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 8(4):289–327, 1973.

[59] Dan E Miller, Robert A Hintz, and Carl J Couch. The elements and structure of openings. *Sociological Quarterly*, 16(4):479–499, 1975.

[60] Graham Button. On varieties of closings. *Interaction competence*, 93:148, 1990.

[61] Jeffrey David Robinson. Getting down to business: Talk, gaze, and body orientation during openings of doctor-patient consultations. *Human Communication Research*, 25(1):97–123, 1998.

[62] Curtis D LeBaron and Stanley E Jones. Closing up closings: Showing the relevance of the social and material surround to the completion of interaction. *Journal of Communication*, 52(3):542–565, 2002.

[63] Kris M Markman. "so what shall we talk about" openings and closings in chat-based virtual meetings. *The Journal of Business Communication (1973)*, 46(1):150–170, 2009.

[64] Hiroko Itakura. Describing conversational dominance. *Journal of Pragmatics*, 33(12):1859–1880, 2001.

[65] Annette L Stanton, Sharon Danoff-Burg, Christine L Cameron, Michelle Bishop, Charlotte A Collins, Sarah B Kirk, Lisa A Sworowski, and Robert Twillman. Emotionally expressive coping predicts psychological and physical adjustment to breast cancer. *Journal of Consulting and Clinical Psychology*, 68(5):875, 2000.

[66] Amy M Sullivan, Matthew D Lakoma, Robin K Matsuyama, Laurie Rosenblatt, Robert M Arnold, and Susan D Block. Diagnosing and discussing imminent death in the hospital: a secondary analysis of physician interviews. *Journal of Palliative Medicine*, 10(4):882–893, 2007.

[67] Lauren A McCormack, Katherine Treiman, Douglas Rupert, Pamela Williams-Piehota, Eric Nadler, Neeraj K Arora, William Lawrence, and Richard L Street Jr. Measuring patient-centered communication in cancer care: a literature review and the development of a systematic approach. *Social Science & Medicine*, 72(7):1085–1095, 2011.

[68] Ronald M Epstein, Brian S Alper, and Timothy E Quill. Communicating evidence for participatory decision making. *JAMA*, 291(19):2359–2366, 2004.

[69] Justin Chow and Helen Senderovich. It's time to talk: challenges in providing integrated palliative care in advanced congestive heart failure. a narrative review. *Current Cardiology Reviews*, 14(2):128–137, 2018.

[70] Barbara Kimbell, Kirsty Boyd, Marilyn Kendall, John Iredale, and Scott A Murray. Managing uncertainty in advanced liver disease: a qualitative, multiperspective, serial interview study. *BMJ open*, 5(11), 2015.

[71] Robert Gramling, Susan Stanek, Paul KJ Han, Paul Duberstein, Tim E Quill, Jennifer S Temel, Stewart C Alexander, Wendy G Anderson, Susan Ladwig, and Sally A Norton. Distress due to prognostic uncertainty in palliative care: frequency, distribution, and outcomes among hospitalized patients with advanced cancer. *Journal of Palliative Medicine*, 21(3):315–321, 2018.

[72] Laurence A. Clarfeld. Source code for conversation dynamics model analysis. A link to source code will be provided when the article is accepted for publication.

[73] Robert Gramling, Susan Stanek, Susan Ladwig, Elizabeth Gajary-Coots, Jenica Cimino, Wendy Anderson, Sally A Norton, Rebecca A Aslakson, Katherine

Ast, Ronit Elk, et al. Feeling heard and understood: a patient-reported quality measure for the inpatient palliative care setting. *Journal of Pain and Symptom Management*, 51(2):150–154, 2016.

[74] MACRA Palliative Care Quality Measures Project, 2019.

[75] NEJM Catalyst. What is patient-centered care? *NEJM Catalyst*, 3(1), 2017.

[76] Alastair Baker. *Crossing the quality chasm: a new health system for the 21st century*, volume 323. British Medical Journal Publishing Group, 2001.

[77] Ronald Epstein. *Patient-centered communication in cancer care: promoting healing and reducing suffering*. Number 7. US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, 2007.

[78] Luke T Ingersoll, Stewart C Alexander, Jeff Priest, Susan Ladwig, Wendy Anderson, Kevin Fiscella, Ronald M Epstein, Sally A Norton, and Robert Gramling. Racial/ethnic differences in prognosis communication during initial inpatient palliative care consultations among people with advanced cancer. *Patient Education and Counseling*, 102(6):1098–1103, 2019.

[79] Karan R Chhabra, Kathryn I Pollak, Stephanie J Lee, Anthony L Back, Roberta E Goldman, and James A Tulsky. Physician communication styles in initial consultations for hematological cancer. *Patient Education and Counseling*, 93(3):573–578, 2013.

[80] JA Gómez-García, Laureano Moro-Velázquez, and Juan Ignacio Godino-Llorente. On the design of automatic voice condition analysis systems. part i: Review of

concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199, 2019.

[81] JA Gómez-García, Laureano Moro-Velázquez, and Juan Ignacio Godino-Llorente. On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors. *Biomedical Signal Processing and Control*, 48:128–143, 2019.

[82] Björn W Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

[83] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421, 2019.

[84] Firoj Alam, Morena Danieli, and Giuseppe Riccardi. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61, 2018.

[85] Zhuohao Chen, James Gibson, Ming-Chang Chiu, Qiaohong Hu, Tara K Knight, Daniella Meeker, James A Tulsky, Kathryn I Pollak, and Shrikanth Narayanan. Automated empathy detection for oncology encounters. *arXiv preprint arXiv:2007.00809*, 2020.

[86] Hiroyuki Sogo. Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Research Methods*, 45(3):684–695, 2013.

[87] Shima Tabibian. A survey on structured discriminative spoken keyword spotting. *Artificial Intelligence Review*, 53(4):2483–2520, 2020.

## 4.6 Supplemental Figures and Tables

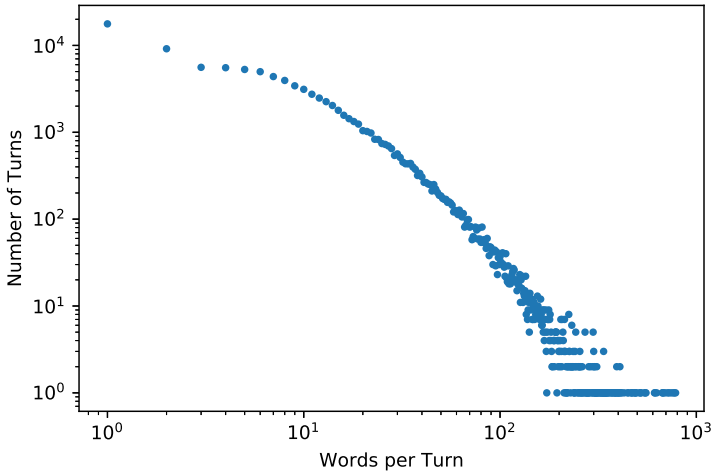



*Figure 4.9:* ***Words per turn in PCCRI corpus.*** *The number of words per turn, for each of the 101,563 turns in the PCCRI corpus.*

*Table 4.2:* ***Classification using CODYMs from emotional content in PCCRI corpus.*** *Classification accuracy of trained Random Forests for predicting which conversations contained at least one instance of anger or fear. Classifiers were trained using transition frequencies of each conversation as input features, for CODYMs of orders one through five. For each model order, we show the number of input features (# feat), and the mean ($\mu$) and the standard deviation ($\sigma$) of the % classification accuracy, averaged over 1000 trained classifiers, both for data unstratified by speaker type, and for stratified data (where CODYMs were populated by patient and clinician turns, separately, for each conversation). P-values are calculated from Z-scores based on $\mu$ and $\sigma$, relative to the null hypothesis that the prediction accuracy is not better than random (i.e., $\leq 50\%$).*

| | Unstratified Data | | | | Stratified Data | | | |
|---|---|---|---|---|---|---|---|---|
| **Order** | **# feat.** | $\mu$ (%) | $\sigma$ (%) | $p$ | **# feat.** | $\mu$ (%) | $\sigma$ (%) | $p$ |
| 1 | 4 | 58.8 | 5.3 | 0.048 | 8 | 59.5 | 5.5 | 0.042 |
| 2 | 8 | 60.3 | 5.2 | 0.024 | 16 | 62.4 | 5.5 | 0.012 |
| 3 | 16 | 64.4 | 5.5 | 0.004 | 32 | 64.5 | 5.3 | 0.003 |
| 4 | 32 | 64.2 | 5.2 | 0.003 | 64 | 64.8 | 5.6 | 0.004 |
| 5 | 64 | 65.4 | 5.3 | 0.002 | 128 | 67.3 | 5.2 | $< 0.001$ |

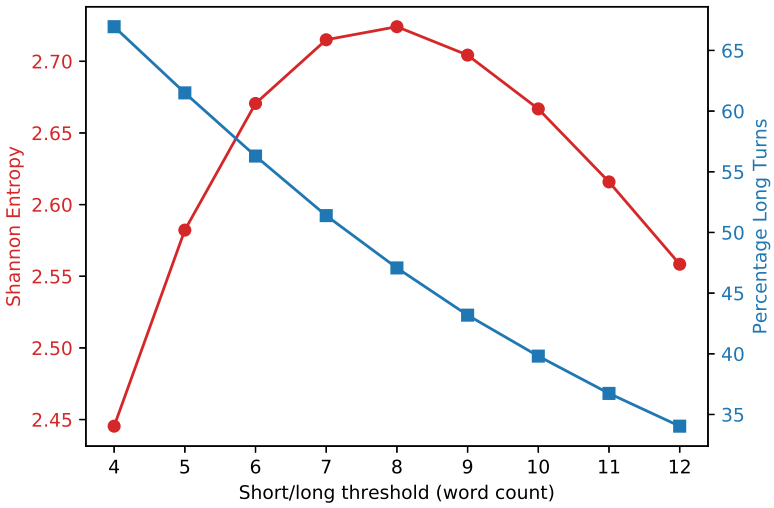



*Figure 4.10:* ***Binarization of turn length in PCCRI corpus.*** *Shannon entropy (information content) of transitions (red curve through circles, left y-axis) and the percentage of long turns for varying short/long thresholds (blue curve through squares, right y-axis) in a $3^{rd}$-order CODYM of the PCCRI corpus. Shannon Entropy is calculated $S = \sum_i f_i \log f_i$ for the frequency $f_i$ of each transition. The short/long threshold is defined such that for a threshold, $t$, any turn with $t$ or more words is considered long. For all experiments in this study, we define short turns to be 7 or fewer words and long turns to be 8 or more words.*

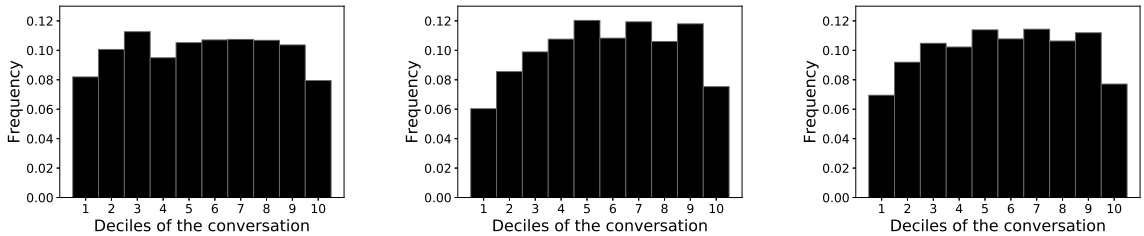



*Figure 4.11:* ***Temporal use of hedging terms in PCCRI corpus.*** *Temporal distribution of all turns in the PCCRI corpus that include hedging terms for (a) patients, (b) clinicians, and (c) all speakers across narrative time.*

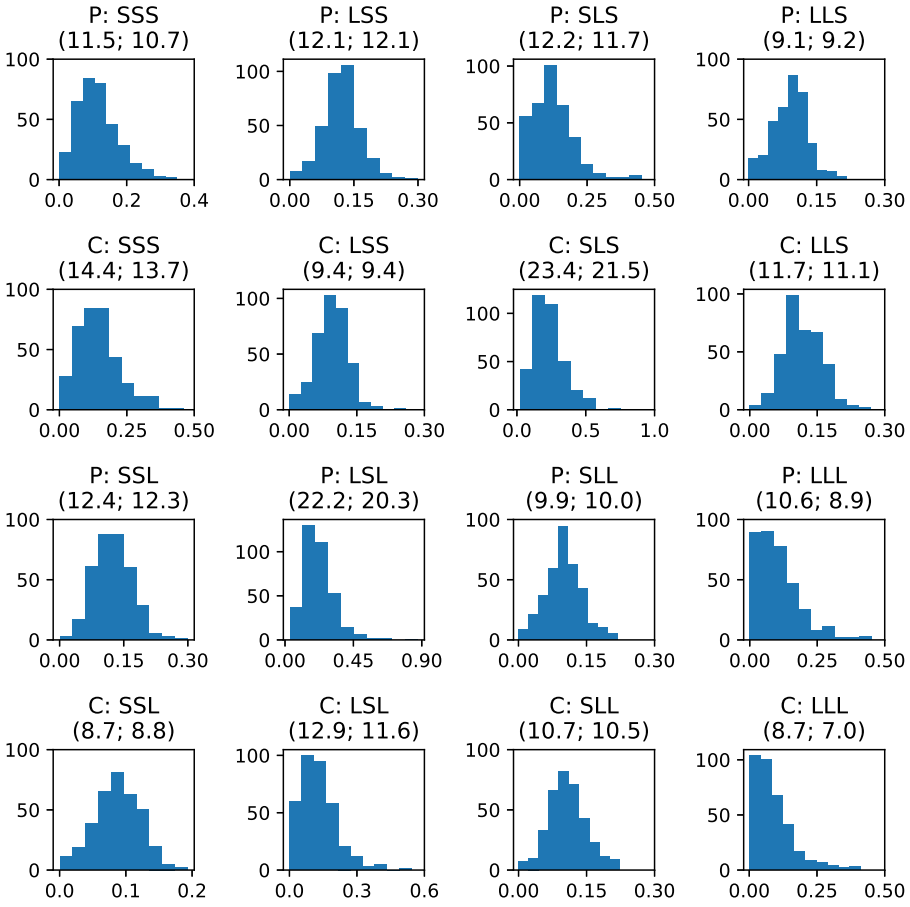



*Figure 4.12:* ***State distributions in PCCRI corpus.*** *The distribution of each state in a $3^{rd}$-order CODYM, stratified by patient and clinician turns, across all 355 conversations in the PCCRI corpus. Each distribution is labeled by patient (P) or clinician (C) turns, the state, and parenthetically the mean and median values, in that order.*

*Table 4.3:* ***Hedging terms.*** *All "hedging terms" (from [51]) along with the number of times each term was used by patients and clinicians in the PCCRI corpus. Note that terms ending in a "*" represent roots and match any word that begins with this term. In addition, "allude to" is a 2-word term.*

| **Term** | **Patient** | **Clinician** |
|---|---|---|
| think | 2652 | 4204 |
| guess | 453 | 199 |
| worr* | 235 | 316 |
| hope | 162 | 293 |
| expect | 79 | 163 |
| seem | 78 | 159 |
| imagine | 41 | 108 |
| consider | 30 | 90 |
| risk | 21 | 115 |
| doubt | 18 | 8 |
| suppose | 16 | 7 |
| predict | 9 | 24 |
| suggest | 7 | 59 |
| theor* | 7 | 9 |
| estimate | 5 | 14 |
| foresee | 3 | 4 |
| anticipate | 2 | 42 |
| assess | 2 | 17 |
| perceive | 2 | 4 |
| contemplate | 1 | 2 |
| presume | 1 | 1 |
| ponder | 1 | 0 |
| suspect | 0 | 18 |
| hint | 0 | 2 |
| speculate | 0 | 1 |
| imply | 0 | 1 |
| hypothesi* | 0 | 1 |
| prognosticate | 0 | 0 |
| presuppose | 0 | 0 |
| postulate | 0 | 0 |
| misjudge | 0 | 0 |
| misinterpret | 0 | 0 |
| infer | 0 | 0 |
| deem | 0 | 0 |
| allude to | 0 | 0 |

*Table 4.4:* ***Treatment terms.*** *All "treatment terms" considered (from [45]), with the number of times each term was used by patients and clinicians in the PCCRI corpus.*

| **Term** | **Patient** | **Clinician** | | **Term** | **Patient** | **Clinician** |
|---|---|---|---|---|---|---|
| hospice | 413 | 1435 | | fentanyl | 64 | 138 |
| chemo | 369 | 157 | | methadone | 63 | 200 |
| radiation | 303 | 467 | | line | 63 | 66 |
| medicine | 289 | 816 | | chemotherapy | 58 | 350 |
| surgery | 236 | 199 | | machine | 56 | 126 |
| medication | 195 | 459 | | treatments | 54 | 469 |
| dilaudid | 186 | 363 | | ativan | 48 | 67 |
| treatment | 181 | 586 | | milligrams | 47 | 90 |
| pills | 161 | 177 | | liquid | 47 | 76 |
| tube | 121 | 380 | | button | 45 | 116 |
| pill | 120 | 114 | | medicines | 43 | 402 |
| morphine | 111 | 212 | | ventilator | 42 | 159 |
| medical | 107 | 455 | | trial | 40 | 74 |
| oxycodone | 98 | 192 | | feeding | 40 | 67 |
| therapy | 96 | 166 | | management | 39 | 217 |
| iv | 93 | 304 | | treat | 38 | 191 |
| tylenol | 85 | 83 | | icu | 37 | 70 |
| drug | 83 | 45 | | fluids | 34 | 83 |
| oxygen | 80 | 81 | | antibiotics | 33 | 132 |
| drugs | 80 | 38 | | oral | 21 | 82 |
| medications | 79 | 425 | | nutrition | 20 | 107 |
| dialysis | 78 | 154 | | doses | 17 | 135 |
| patch | 73 | 214 | | resuscitation | 17 | 101 |
| procedure | 72 | 75 | | cpr | 17 | 78 |
| dose | 66 | 430 | | milligram | 9 | 33 |
| meds | 65 | 46 | | | | |

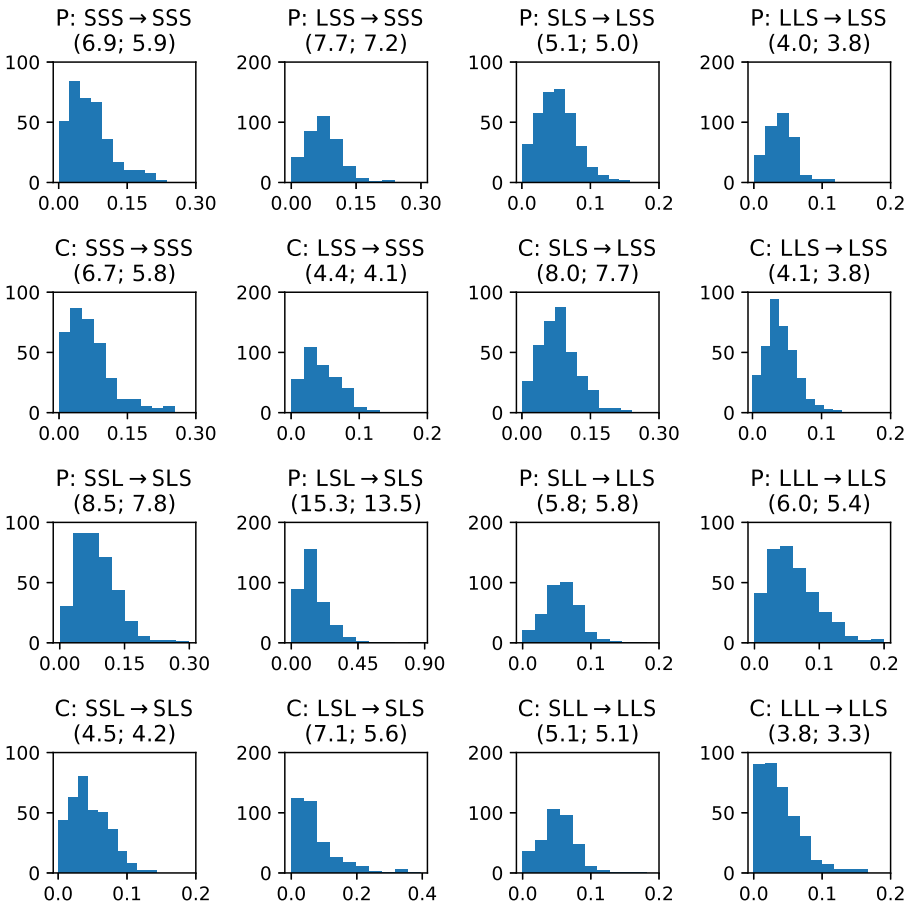



*Figure 4.13:* ***Transition distributions of short turns in PCCRI corpus.*** *The distribution of frequencies on each* ***short*** *transition in 3$^{rd}$-order CODYMs, stratified by patient and clinician turns, across all 355 conversations in the PCCRI corpus. Each distribution is labeled by patient (P) or clinician (C) turns, the transition, and parenthetically the mean and median values, in that order.*

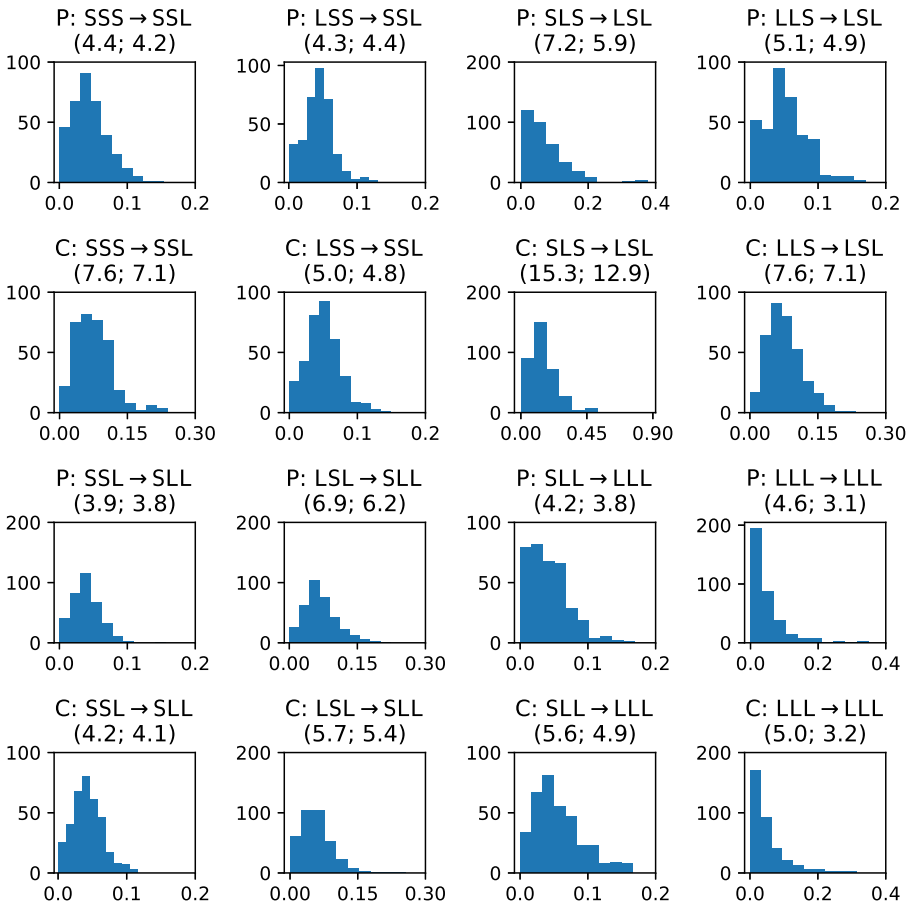



*Figure 4.14:* ***Transition distributions of long turns in PCCRI corpus.*** *The distribution of frequencies on each* ***long*** *transition in $3^{rd}$-order CODYMs, stratified by patient and clinician turns, across all 355 conversations in the PCCRI corpus. Each distribution is labeled by patient (P) or clinician (C) turns, the transition, and parenthetically the mean/median values.*

*Table 4.5:* ***Significance tests of state distributions in PCCRI corpus.*** *P-values comparing the state distributions of 3rd-order CODYMs in the PCCRI corpus, stratified by patient and clinician (shown in Fig. 4.12), of observed patient* vs. *observed clinician, observed patient* vs. *null patient, and observed clinician* vs. *null clinician models, using a 2-sample Kolmogorov–Smirnov Test.*

| State | Patient *vs.* Clinician | Patient *vs.* Null | Clinician *vs.* Null |
|---|---|---|---|
| SSS | <0.001 | <0.001 | <0.001 |
| SSL | <0.001 | <0.001 | <0.001 |
| SLS | <0.001 | <0.001 | <0.001 |
| SLL | <0.001 | <0.001 | <0.001 |
| LSS | <0.001 | <0.001 | <0.001 |
| LSL | <0.001 | <0.001 | <0.001 |
| LLS | 0.063 | <0.001 | <0.001 |
| LLL | 0.002 | <0.001 | <0.001 |

*Table 4.6:* ***Significance tests of transition distributions in PCCRI corpus.*** *P-values comparing the transition distributions of 3$^{rd}$-order CODYMs, stratified by patient and clinician (shown in Figs. 4.13 and 4.14), of observed patient* vs. *observed clinician, observed patient* vs. *null patient, and observed clinician* vs. *null clinician models using a 2-sample Kolmogorov–Smirnov Test.*

| **Transition** | **Patient *vs.* Clinician** | **Patient *vs.* Null** | **Clinician *vs.* Null** |
|---|---|---|---|
| SSS$\xrightarrow{S}$SSS | 0.811 | <0.001 | <0.001 |
| SSS$\xrightarrow{L}$SSL | <0.001 | <0.001 | <0.001 |
| LSS$\xrightarrow{S}$SSS | <0.001 | <0.001 | <0.001 |
| LSS$\xrightarrow{L}$SSL | <0.001 | <0.001 | <0.001 |
| SLS$\xrightarrow{S}$LSS | <0.001 | <0.001 | <0.001 |
| SLS$\xrightarrow{L}$LSL | <0.001 | <0.001 | <0.001 |
| LLS$\xrightarrow{S}$LSS | 0.504 | <0.001 | <0.001 |
| LLS$\xrightarrow{L}$LSL | <0.001 | <0.001 | <0.001 |
| SSL$\xrightarrow{S}$SLS | <0.001 | <0.001 | <0.001 |
| SSL$\xrightarrow{L}$SLL | 0.187 | <0.001 | <0.001 |
| LSL$\xrightarrow{S}$SLS | <0.001 | <0.001 | <0.001 |
| LSL$\xrightarrow{L}$SLL | <0.001 | <0.001 | <0.001 |
| SLL$\xrightarrow{S}$LLS | 0.014 | <0.001 | <0.001 |
| SLL$\xrightarrow{L}$LLL | <0.001 | <0.001 | <0.001 |
| LLL$\xrightarrow{S}$LLS | <0.001 | <0.001 | <0.001 |
| LLL$\xrightarrow{L}$LLL | 0.627 | <0.001 | <0.001 |

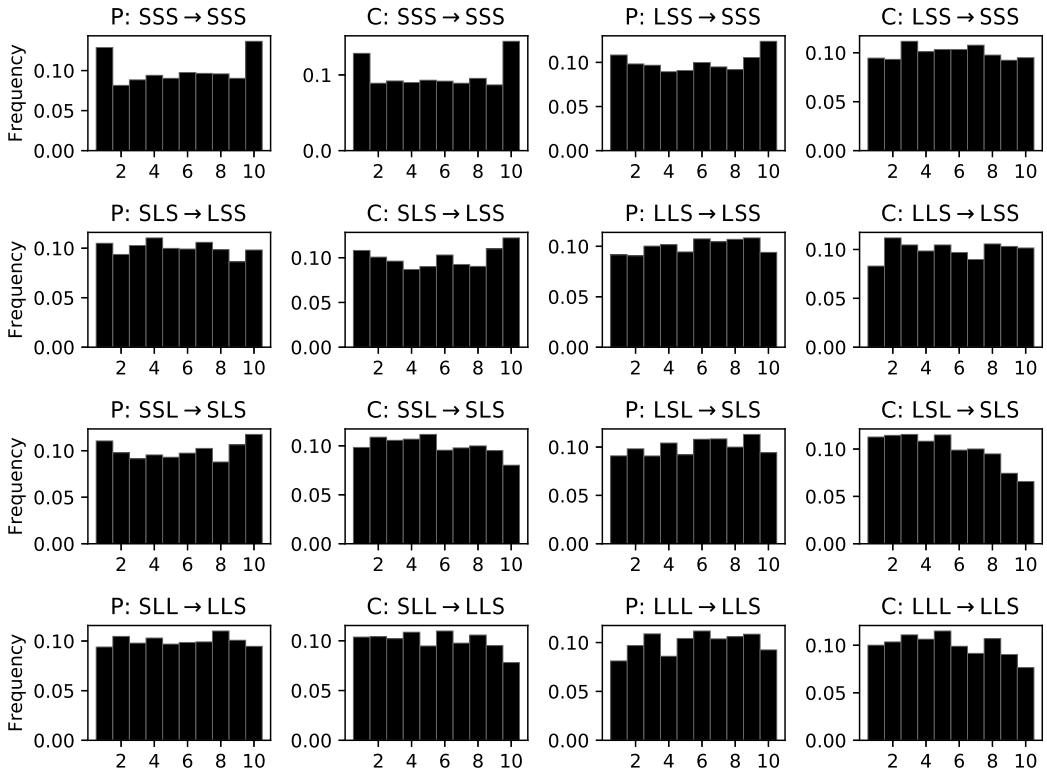



*Figure 4.15:* ***Temporal changes in transitions of short turns in PCCRI corpus.*** *Histograms of transition frequencies of all* ***short*** *turns in $3^{rd}$-order CODYMs over 10 conversational deciles (normalized, such that the sum of all bins is 1.0), stratified by the patient and clinician turns in the PCCRI corpus.*

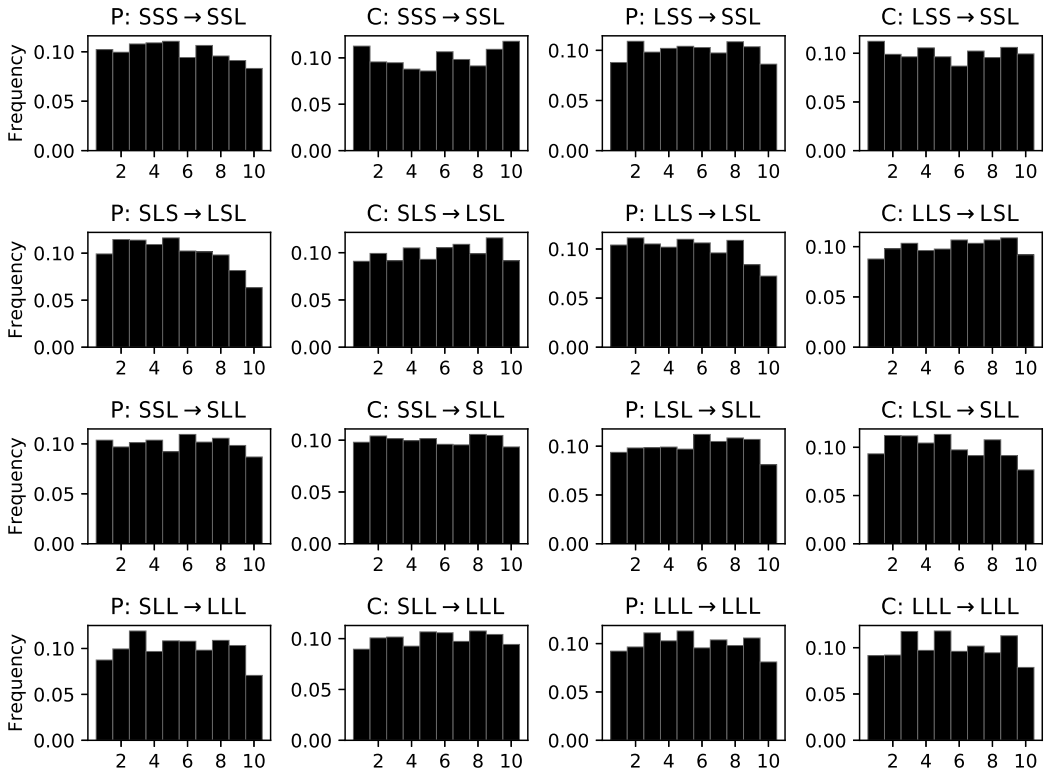



*Figure 4.16:* ***Temporal changes in transitions of long turns in PCCRI corpus.*** *Histograms of transition frequencies of all* ***long*** *turns in $3^{rd}$-order CODYMs over 10 conversational deciles (normalized, such that the sum of all bins is 1.0), stratified by the patient and clinician turns.*

*Table 4.7:* ***Complete word clusters from CODYMs of PCCRI corpus.*** *All words appearing in each of the six unsupervised clusters of 114 words, based on similarities in transition frequencies in a $2^{nd}$-order CODYM. Words are parenthetically followed by the number of times each occurred in the PCCRI corpus. To comply with HIPPA privacy protection, all names and dates were anonymized; "p_name" refers to the name of any patient or family member, "u_name" refers to the name of an unknown speaker, and "u_date'' refers to any specific date.*

| **Cluster Moniker (# words)** | **Unique Words (# occurrences in PCCRI corpus)** |
|---|---|
| **Strong Continuers (4 words)** | hm (1987); hmm (1896); mm (1263); aha (277) |
| **Moderate Continuers (10 words)** | yeah (15367); okay (12332); oh (3607); uh (3288); yes (2611); huh (527); absolutely (402); wow (354); yup (334); yep (303) |
| **Weak Continuers (19 words)** | right (9663); no (6009); p_name (2578); nice (1032); great (775); fine (679); sorry (630); exactly (465); alright (438); meeting (266); true (247); appreciate (210); correct (191); please (188); perfect (150); appetite (141);beautiful (129);u_date (118); boy (114) |
| **Openers/Closers (7 words)** | thank (1232); meet (565); hi (404); thanks (192); welcome (135); hello (125); bye (111) |
| **Clinical Talk (51 words)** | um (5993); also (1350); sort (978); may (907); comfort (685); symptoms (623); treatments (525); whether (423); chemotherapy (409); continue (337); disease (267); methadone (263); focus (256); lungs (251); means (233); treat (230); acting (228); system (194); symptom (186); machine (183); oncology (182); likely (176); safe (170); fix (163); recommend (159); show (158); illness (157); page (148); folks (145); decided (139); controlled (135); outpatient (133); homes (130); setting (126); clinic (125); further (124); belly (124); approach (124); death (122); spot (120); consider (120); kept (117); ended (114); words (109); example (108); icu (106); cpr (105); finally (105); depending (103); depressed (102); stronger (101) |
| **Potpourri (23 words)** | well (5239); old (263); 5 (204); u_name (178);hurts (176); tylenol (170); totally (169); agree (165); friend (165); button (159); lives (144); middle (142); concern (141); card (138); risk (137); cold (117); covered (116); fair (116); hit (115); confused (107); scary (104); mg (104); funny (103) |

*Table 4.8:* ***Mean transition frequencies in word clusters of PCCRI corpus.*** *Mean transition frequencies (%) for $2^{nd}$-order CODYMs for all words in the PCCRI corpus (ALL), and for the words in each of the six clusters: Strong Continuers (SC), Moderate Continuers (MC), Weak Continuers (WC), Openers/Closers (OC), Clinical Talk (CT), and PotPourri (PP).*

| Transition | ALL | SC | MC | WC | OC | CT | PP |
|---|---|---|---|---|---|---|---|
| SS$\stackrel{S}{\rightarrow}$SS | 3.0 | 13.3 | 14.2 | 11.0 | 27.6 | 1.3 | 4.4 |
| LS$\stackrel{S}{\rightarrow}$SS | 2.7 | 7.4 | 8.5 | 7.0 | 12.7 | 1.3 | 4.0 |
| SL$\stackrel{S}{\rightarrow}$LS | 3.3 | 50.9 | 26.7 | 12.1 | 11.1 | 1.1 | 3.9 |
| LL$\stackrel{S}{\rightarrow}$LS | 2.0 | 17.6 | 13.8 | 6.3 | 4.4 | 0.9 | 2.9 |
| SS$\stackrel{L}{\rightarrow}$SL | 18.1 | 2.0 | 8.5 | 15.5 | 17.1 | 15.9 | 20.0 |
| LS$\stackrel{L}{\rightarrow}$SL | 35.8 | 2.0 | 11.1 | 21.4 | 12.5 | 47.7 | 24.5 |
| SL$\stackrel{L}{\rightarrow}$LL | 17.7 | 4.0 | 10.0 | 13.7 | 7.7 | 15.1 | 19.5 |
| LL$\stackrel{L}{\rightarrow}$LL | 17.4 | 2.9 | 7.2 | 12.9 | 6.9 | 16.8 | 20.8 |

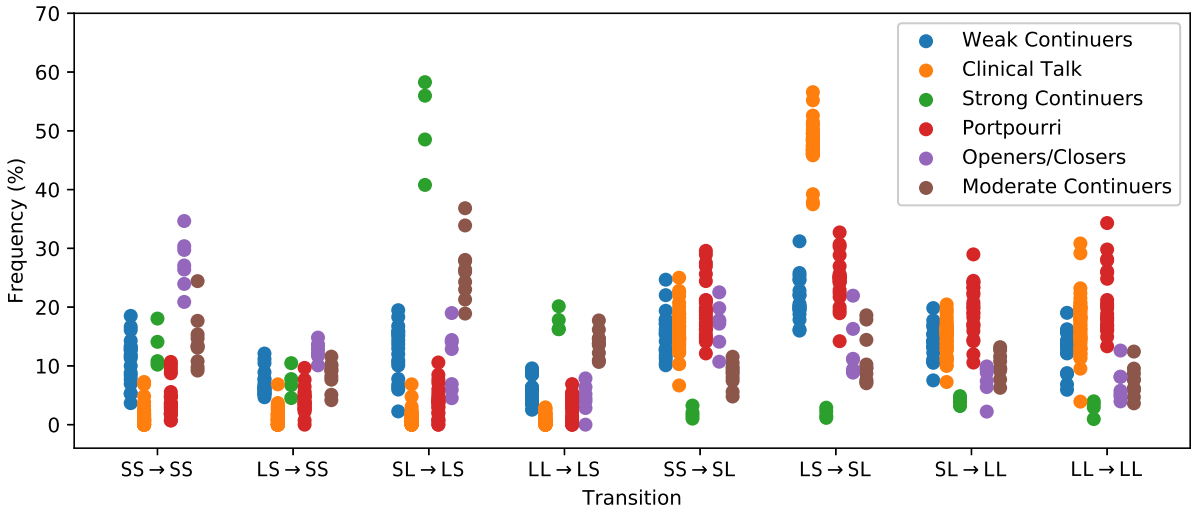



*Figure 4.17:* ***Distributions of transition frequencies in word clusters of PCCRI corpus.*** *Transition frequencies, by word cluster, for all words in the PCCRI coprus that are used 100 or more times, and whose transition frequencies differ by at least 10% from expected. The ordering of clusters (from left to right) in the figure match the ordering of cluster names (from top to bottom) in the legend.*

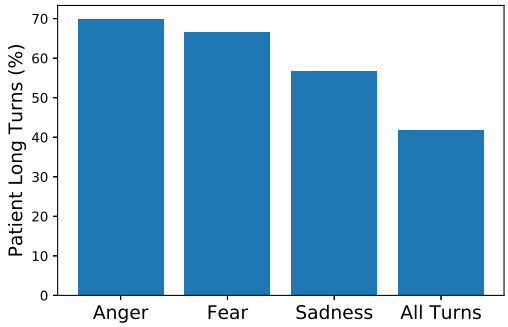



*Figure 4.18:* ***Turn length by emotion in PCCRI corpus.*** *Percentage of turns in which distressing emotion (anger, fear, sadness) are expressed that are long, compared to all patient turns.*

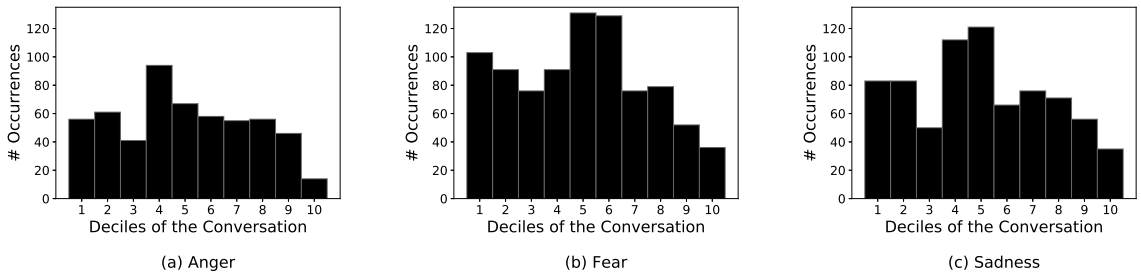



*Figure 4.19:* ***Temporal expressions of emotion in PCCRI corpus.*** *Temporal distribution of turns in the PCCRI corpus that include patient expressions of (a) anger, (b) fear, and (c) sadness across narrative time.*

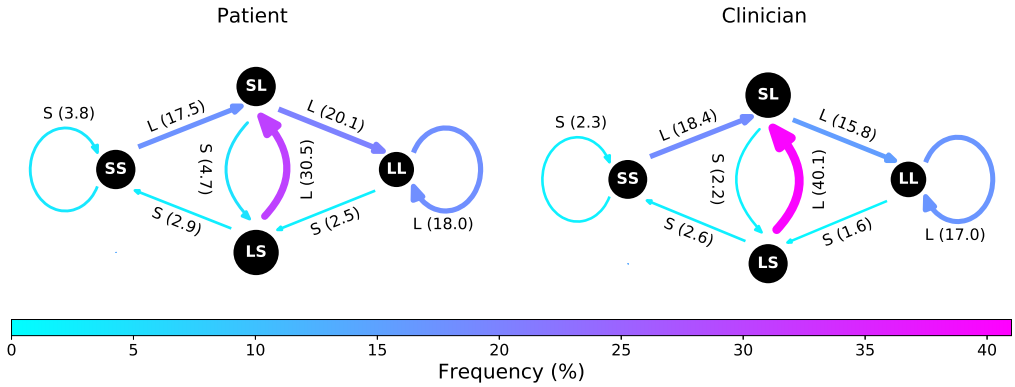



*Figure 4.20:* ***CODYMs of word usage in stratified PCCRI corpus.*** *State and transition frequencies of all words in the PCCRI corpus, stratified by patient and clinician turns. Transition labels indicate the length of the turn on the transition, parenthetically followed by the percentage of word occurrence on that transition. Edge thickness and color indicate %Observed for each transition, and node diameter indicates %Observed for each state. 1000 null models were created according to these frequencies to determine expected values for comparison to frequencies of hedging or treatment terms.*

# Chapter 5

# Conclusions

## 5.1 Summary of Contributions

The primary, original contributions of this dissertation are organized by chapter and bulleted here:

### 5.1.1 Chapter 2

- Proposed copula analysis for calculating risk given spatially correlated outages.
- Defined a semi-metric for determining the distance between two line segments that conforms to expectations in the context of a common-cause event threatening an electric transmission system.
- Demonstrated risk analysis on a larger (>3× from previous work) synthetic power systems test case that included more realistic topography.
- Developed novel methods for identifying lower- and upper-bound estimates of

defective 3-sets ($N - 3$ malignancies) in large networks, such as used here.

- Using aforementioned methods, demonstrated that:
  - Incorporating spatial correlation between component outages into risk estimation of cascading failure can lead to a substantial increase in risk.
  - Under the assumption of correlated component outages, higher-order defective sets contribute proportionally more to overall risk.

### 5.1.2 Chapter 3

- Introduced SIGHT, a novel deterministic group-testing algorithm for identifying defective hyper-edges (sets) in a hypergraph where:
  - Defective sets are too numerous to tractably enumerate.
  - Defective sets are variably sized, but with a known lower-bound.
  - An upper-bound size of defective sets to be found can be specified.
  - Tests can produce false negatives.
- Created a domain-independent, test problem generator with variable numbers of defective $k$-sets for $k > 2$, an unknown number of false negatives ($k$-sets that test as negative but actually do contain a minimal defective $k$-set), and a tunable negative:positive test cost ratio.
- Showed that Random Chemistry (RC) is more efficient in applications where the positive:negative cost ratio of tests is high (due to a lower positive:negative test use ratio than SIGHT), such as the power systems application from Chapter 2,

but that SIGHT is expected to outperform RC in applications where the cost ratio is low.

- When the defective:non-defective cost ratio is high, group-testing algorithms are best parameterized to have a low initial set size ($a_0$), resulting in a high initial failure rate, contrary to previosly adopted heuristics. We find that for the Western US test case, reducing $a_0$ from 320 to 96 acheives a nearly 2× speedup.

### 5.1.3 Chapter 4

- Introduced CODYM analysis, a novel, Markov-based method for quantifying, visualizing, and interpreting information flow patterns in conversations. The model is uniquely characterized by:
  - Markov states that are defined using speaker turns of variable length rather than a fixed interval of time.
  - a context-independent model that preserves the privacy of conversational participants.
  - ease of automation.
  - semantically interpretable visualizations for understanding information flow patterns, including one-way exchanges and short/long two-way exchanges.
- Used unsupervised word clustering to reveal groups of words that have specialized usage related to their information sharing patterns, thus validating that CODYMs capture semantically meaningful patterns related to:

  - Conversational "openers" and "closers" that share similar patterns, over-represented in short two-way exchanges.
  - Groups of "continuers" that appear frequently in the short turns of one-way exchanges, naturally stratified into three clusters by the degree of specialization of the words in each cluster.
  - "Clinical talk" that occurs most frequently during the long turns of one-way exchanges, especially when the clinician is speaking.
  - A "portpouri" cluster that was less well-defined, but contained some words related to feelings/risk that occur more frequently on long one-way exchanges.

- Revealed conversation-wide normative patterns of serious illness conversations:
  - One-way information flow is over-represented for both patients and clinicians and depressed for all other information sharing patterns, relative to what is expected by chance.
  - One-way flow is more often from clinician-to-patient than from patient-to-clinician, as expected given the institutional setting and pre-defined roles of conversational participants.

- Revealed dynamic changes in normative patterns over the course of serious illness conversations:
  - Short two-way exchanges are most numerous at the beginning and ends of conversations, corresponding to conversational openings and closings.

  - One-way information flow from patient-to-clinician (but not from clinician-to-patient) decreases through the second half of the conversation, a possible indication that the patients have sufficiently expressed themselves as the conversations naturally wind down to a close (a marker of patient-centered care).

- Examined information flow patterns of:
  - Treatment terms, which are used frequently by clinicians in the long turns of one-way exchanges, but not by patients, indicating a distinction in how these terms are used between patient and clinician.
  - Hedging terms, which are used more frequently in long two-way exchanges, by both patients and clinicians, indicating a possible bilateral sharing regarding uncertainty about difficult topics (a marker of patient-centered care).

- Showed that patient expressions of distressing emotion are associated with deviations from the normative information flow patterns (in ways that are consistent with patient-centered care):
  - Locally (during the turns in which emotion is expressed), the dominant pattern of one-way information flow flips from clinician-to-patient to patient-to-clinician.
  - Globally (comparing conversations with expressed emotion to those without), long two-way exchanges and one-way exchanges from patient-to-clinician occur more frequently, while short two-way exchanges and one-way exchanges from clinician-to-patient occur less frequently.

## 5.2 Future Work and New Preliminary Findings

### 5.2.1 Chapter 2

This dissertation proposed several new methodologies using "hidden networks" to address challenging problems in disparate fields of study, but additional work is needed to both fully apply and validate these methods. We presented a copula-based method for quantifying risk to the electric transmission system from cascading failures, given correlated outages, but have thus far only applied this method to grid models that are either synthetic or lacking in topographical features. While most real-world data from actual transmission systems are typically not available to academic researchers due to security restrictions, acquiring data about the correlation of component failures in order to make better informed choices about which correlation and copula functions may be a good next step towards applying these methods in a practical setting.

As we consider larger, more complex infrastructures, the number of components (and defective sets) is sure to increase, making it more challenging to predict the number of defective sets. Because it is intractable to enumerate all defective sets, it is not feasible to validate estimation methods from cascading simulations. This necessitates the development of a new test problem in which group testing can be applied but where the number of defective sets can be determined *a priori* without the need to fully enumerate them. Specifically, a desired test case would have the following properties:

1. There exists a function that can determine if an arbitrarily large set has any defective subsets.

2. The number of defective sets, for $k > 3$, may exceed $10^7$.

3. The exact number of defective $k$-sets can be derived efficiently.

4. There can be orders of magnitude more defective $k+1$-sets than defective $k$-sets.

5. A defective $k$-set can be arbitrarily large.

6. The frequency with which elements appear in defective sets is heavy-tailed.

In order to mimic the observed trends in the distribution of defective sets in the power systems application, a test case should be capable of generating defective sets of varied size, where the number of defective $k$-sets are large and increase by an order of magnitude with $k$ (as has been observed in the Polish and Western US test cases). Furthermore, in the real power grid a small number of components are typcially responsible for a large proportion of the risk associated with the system, which must be reflected in our test case in order to generate the heterogeneous sampling probabilities that result when appying group-testing to the problem. Our test function should have 100% reliability, though some small rate of error can be stochastically added to simulate the false-negatives that may occur in practice.

Consider a universal set of elements $V = \{i \in \mathbb{Z}^+ \mid i \leq N\}$ and a special value $\alpha$. I define a defective set to be one $S \subset V$ such that if the product of the elements in $S$ are divisible by $\alpha$, then the set is defective. Based on this definition, our "test" function is trivially defined to return 1 when a (non-minimal) defective set is found (and 0 otherwise):

$$f(S) = \begin{cases} 1 & \text{if } (\prod S) \mid \alpha = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

This definition of defective sets meets most of the criteria above, except it is not immediately clear how to efficiently calculate the number of such sets that exist for a given $N$ and $\alpha$. I conjecture that for the general case, there is no efficient solution, however by imposing certain restrictions on our choice of $\alpha$, I have found efficient solutions that may be suitable for further study of set-size estimation methods. For example, when $\alpha = p_1 p_2 p_3 p_4$ where each $p_i$ is a unique prime number, a linear-time algorithm exists for quantifying the number of defective sets. This special case will result in defective sets up to size 4. It is yet to be determine if such a test case is sufficient for assessing algorithms for predicting set size, or whether test cases with larger defective sets are needed.

Another important step in improving set size estimation is to better understand the heterogeneity of group testing algorithms when used to collect a large sample of defective sets. It has been noted in this work that defective sets with 'risky' branches (that appear frequently in defective sets) are under-sampled by group testing algorithms. In fact, when the total set of malignancies is known, the sampling bias introduced by group testing algorithms can be directly calculated.

Consider the minimal example of a 4-bus network with three $N-2$ malignancies, namely $\Omega_2 = \{(1,2),\ (1,3),\ (2,4)\}$. Consider the subset reduction scheme $S = \{4,3,2\}$. The initial four possible subsets of size 3 each have a 25% chance of being selected (Table 5.1, col 1). Each of these subsets contains a malignancy, so RC will be successful on its first attempt. Each subset of 3, however, does not have the same

number of $N-2$ malignancies and may fail on the first attempt. Since RC will re-sample until finding a malignancy, the chances of any $N-2$ being sampled from a given subset of 4 is inversely proportional to the number of $N-2$ malignancies in each subset. The likelihood of each malignancy being chosen from the subsets of 3, along with overall probabilities of being selected, are shown in Table 5.1 cols 3-5.

| Sets of 3 | Sub-sets of 2 | | | Frequency | | |
|---|---|---|---|---|---|---|
| | | | | [1 2] | [1 3] | [2 4] |
| [1 2 3] | [1 2] | [1 3] | [2 3] | $\frac{1}{2}$ | $\frac{1}{2}$ | |
| [1 2 4] | [1 2] | [1 4] | [2 4] | $\frac{1}{2}$ | | $\frac{1}{2}$ |
| [1 3 4] | [1 3] | [1 4] | [3 4] | | 1 | |
| [2 3 4] | [2 3] | [2 4] | [3 4] | | | 1 |
| Total Sampling Probability (%) | | | | 25 | 37.5 | 37.5 |

*Table 5.1: Minimal example of Random Chemistry introducing sampling bias whereby some malignancies are more likely to be sampled than others. Malignancies are highlighted in yellow.*

As has been demonstrated, we can brute-force calculate the sampling bias for small examples of Random Chemistry, for a single reduction step. Calculations akin to those in Table 5.1, however, require enumerating all $\binom{N}{s}$ ways of reducing from a set of size $N$ to $s$, and all $\binom{N}{s}\binom{s}{k}$ ways of reducing from each set of size $s$ to a set of size $k$. In order to improve upon the efficiency of this calculation, an accurate estimation of bias can be derived by randomly sampling sets of size $s$, although a full description is beyond the scope of this work.

In the simple example where only defective 2-sets are present, we can define a 'Malignancy Graph' where each edge corresponds to a defective set. We have observed in preliminary experimentation that a strong negative correlation exists between the

amount of observed sampling bias of a defective set and the Page Rank centrality of the corresponding edge in the Malignancy Graph. This has led us to interpret sampling bias as a new form of centrality, which we refer to as 'bias centrality', where the most central nodes/edges in the Malignancy Graph are under-sampled by group-testing algorithms. More work is needed to determine additional properties of bias centrality, including the effects on bias of multiple set size reduction steps and higher-order defective $k$-sets ($k > 2$).

### 5.2.2 Chapter 3

In Chapter 3, we have performed a first-of-its-kind comparison between stochastic (Random Chemistry, or RC) and deterministic (SIGHT) group-testing algorithms and described the effects of both the number and relative cost of tests involving defective *vs.* non-defective sets on algorithmic performance, using a domain-independent test problem. In addition to the domain-independent comparison of Chapter 3, in a separate study we compared RC and SIGHT in the application domain of power systems that initially inspired this work [1]. Here, we describe some of the key results of that study.

Our auxiliary study used the same test case of the Western US and same methods for simulating cascades as in Chapter 2. We performed 290,000 paired simulations of RC *vs.* SIGHT with initial set sizes, $a_0$, ranging from 48 to 192. As described previously, each step of the cascade is simulated sequentially, thus for each addition step, additional computational resources are required. It is thus not surprising that tests of defective sets take longer than non-defective sets, as shown in Chapter 3.

In our previous work, the effects of the testing cost ratio on initial set size had not

been considered [2]. Prior convention had assumed that an initial set size should be chosen such that, most of the time, the initial set contains a defective subset, so in the work in Chapter 2 we had set $a_0$ to 320. However, in Chapter 3, we discovered that the optimal $a_0$ varies with the specifics of the problem. A subsequent parameter sweep on the Western US test case revealed that $a_0 = 96$ minimized run times for both RC and SIGHT (Fig. 5.1), even though this increased the initial fail rates to over 97%. This increase in efficiency was due to the reduced chance of more costly 'aborts' (requiring more defective tests) later in the sequence of tests performed by either RC or SIGHT. On the Western US test case, using $a_0 = 96$ rather than $a_0 = 320$, reduced the time per find using RC by almost half. At the optimal $a_0$=96, we found that RC outperformed SIGHT in all of the circumstances considered in the power systems application, even when using a fast DC simulator, despite SIGHT needing fewer overall tests (Fig. 5.1).

We used the observed ratio of defective:non-defective turns per find for each algorithm (for $k_{max} \in \{2, 3, 4\}$) to determine the testing cost ratio required for one algorithm to outperform the other (Fig. 5.2). SIGHT becomes more competitive to RC for smaller $k_{max}$, but the high test cost ratio of DCSIMSEP still favors RC in all scenarios considered. It would be interesting to replicate these experiments using a more sophisticated AC cascade simulator, should one become available that is appropriate for this scale of simulation. As stated in Chapter 3, we expect that a more computationally AC intensive simulator will favor Random Chemistry over SIGHT even more than the results shown here, since the average cost ratio of defective:non-defective tests will be higher.

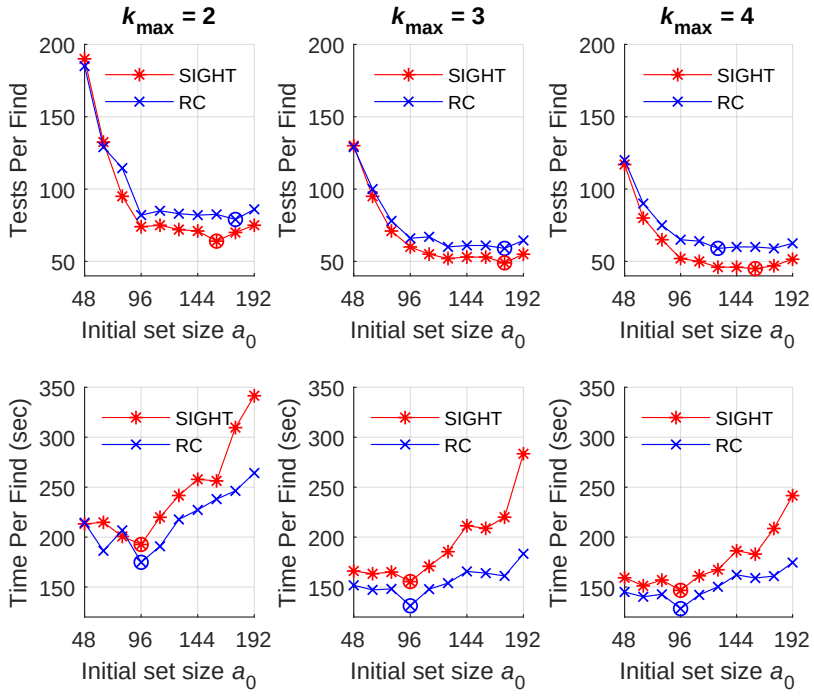



*Figure 5.1: The median number of tests (top row) and median run time, in seconds (bottom row), per defective set found by RC and SIGHT on the Western U.S. power systems test case, for $a_0 \in \{48, 64, 80, 96, 112, 128, 144, 160, 176, 192\}$ stratified by $k_{max} \in \{2, 3, 4\}$ (columns). The cost due to aborted runs in between successful finds is included in these metrics. The empirically minimum data points for each algorithm are circled.*

### 5.2.3 Chapter 4

While the risk estimation methods in power systems have thus far only been applied to synthetic test problems, our novel method for modeling information flow in conversations has been demonstrated on a real-world corpus of serious illness conversations. We demonstrated the method, revealing patterns of information flow in a variety of contexts, but as new conversational features become available, CODYM analysis may inform additional insights into the PCCRI corpus. For example, connectional silences are pauses in a conversation that "feel comforting, affirming, and safe" [3] and have

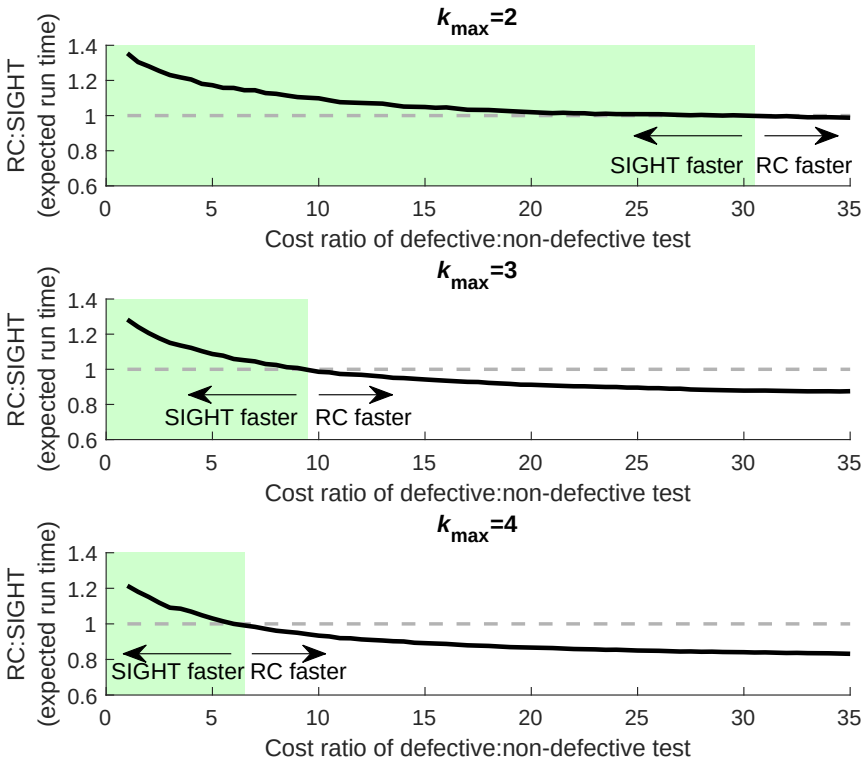



*Figure 5.2: The expected ratio of run time for RC:SIGHT as a function of the ratio of the cost of testing defective sets:non-defective sets. The highlighted regions indicate where SIGHT is expected to be faster than RC. Results are shown for* $k_{max} \in \{2, 3, 4\}$.

been found to have clinical significance in patient communication [4, 5, 6]. Methods for identifying and coding connectional silences in the PCCRI corpus have been recently developed [3, 7] and as these data become available, CODYM analysis could be used to consider the information flow pattern surrounding connectional silences.

In Chapter 4 we stated that it remains to be seen whether the normative patterns observed in the PCCRI corpus capture a conversational "fingerprint" that is unique to serious illness conversations, or whether some of the observed patterns of information flow are ubiquitous across all conversations. Preliminary analysis indicates there

is at least one corpus that breaks with the normative patterns observed in serious illness conversations. On a corpus of transcripts from 553 oral arguments from the US Supreme Court [8], courtesy of Cornell's 'ConvoKit' [9] (which excluded transcripts shorter than 20 speaker turns), the median turn length is 22 words (compared to 8 words in serious illness conversations). If we use a short-long threshold of 22 words, we observe that one-way information flow is under-represented in the observed data compared to what would be expected by chance while short two-way information exchanges, and especially long two-way information exchanges, are over-represented in 3$^{rd}$-order CODYMs (Fig. 5.3).

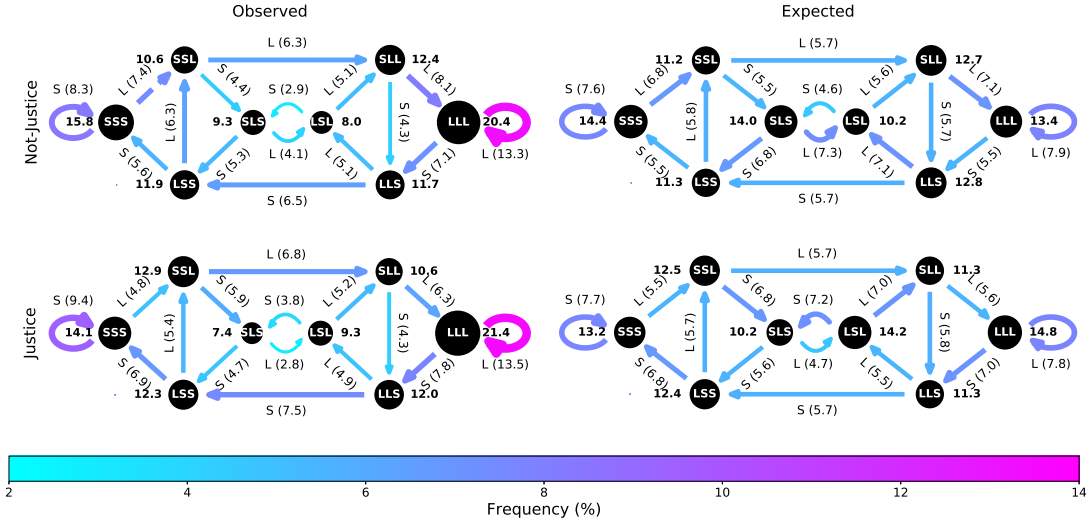



*Figure 5.3: CODYMs of normative patterns of information flow for turns of non-justices (top left) and justices (bottom left), averaged over all conversations in the SCOTUS corpus. Null models were constructed with the same turn length imbalance for non-justices and justices and the same sequential order of non-justice and justice turns in each conversation (right column). Edge thickness and color indicate %Observed for each transition, as shown parenthetically on edge labels. Node diameter indicates %Observed for each state, as shown by the node labels in bold. All state and transition values were significantly different from their corresponding null models, according to the empirically derived 95% confidence intervals, with the exception of transition* $\mathsf{SSS} \overset{L}{\Rightarrow} \mathsf{SSL}$ *for non-justices (indicated by the dashed edge).*

Using median turn length for each corpus effectively normalizes them independently

to account for differences in average turn length between differing corpora. This normalization may be useful for elucidating the patterns that occur within each corpus. If the same threshold (8 words) used in Chapter 4 were used on this corpus, nearly 74% of turns would be long and the resulting CODYM would be less revealing. But when performing other analyses (such as comparing word use), it may be more important to preserve the short/long threshold between corpora. For example, with a threshold of 22 words, short two-way information exchanges have a different semantic interpretation compared to the 8-word threshold used for the PCCRI coprus. Balancing the ability of a fixed threshold to preserve semantic meaning of short/long *vs.* allowing variable thresholds to prevent the washing out of patterns within a corpus will be a challenge that must be addressed before widespread comparisons across corpora can occur.

Like most of the corpora included in Cornell's 'ConoKit' [9], the Supreme Court corpus is rich in metadata. As a demonstration of the versatility of CODYM analysis, further study can compare differences in information flow between individual justices, between justices appointed by Republican *vs.* Democractic presidents, between cases were justices rule in favor of *vs.* against plaintiffs, whether justices dissent from the majority or not, and whether they are male or female. Essentially any conversational feature can be considered in the context of information flow.

While there are a growing number of corpora of conversational transcripts available (for example, see [9, 10, 11, 12, 13]), transcription is a labor-intensive process and may be prohibitive in many circumstances, including cases where the privacy of conversational participants must be preserved. Thus, in Chapter 4, we proposed an alternative formulation of CODYM analysis that uses turn duration (rather than word counts) to define the state space. Using a simple amplitude threshold for the voice

signal of a speaker should permit the identification of the start and end of speaker turns without compromising the anonymity of conversational participants. Identification of the current speaker in a conversation is a native feature of many tele-conferencing services currently in popular use. It is therefore feasible to perform CODYM analysis on any of the thousands of conversations happening remotely on a daily basis. A new startup video conferencing company, InSpace, has been identified as a potential research partner and discussions are underway about incorporating capabilities for CODYM analysis into their platform. The study of remote conversations has taken on new significance in the wake of the Covid-19 global pandemic, where these new technologies have been rapidly adopted without full consideration of how they impact the ways we share information.

In both the power systems and conversation analysis applications, the generalizability of the 'hidden networks' concept has been demonstrated. Additionally, we conclude by offering new avenues of exploration. The universality of networks, and of computing more generally, allow multidisciplinary research, such as this dissertation, to happen. Such collaborations between researchers in different fields will be needed to tackle the increasingly complex challenges that the world will be facing in the years to come.

# Bibliography

[1] Laurence A Clarfeld and Margaret J Eppstein. Group-testing on hypergraphs with variable-cost tests: A power systems case study. *arXiv preprint arXiv:1909.04513*, 2019.

[2] Laurence A Clarfeld, Paul DH Hines, Eric Hernandez, and Margaret J Eppstein. Risk of cascading blackouts given correlated component outages. *IEEE Transactions on Network Science and Engineering*, 2019.

[3] Brigitte N Durieux, Cailin J Gramling, Viktoria Manukyan, Margaret J Eppstein, Donna M Rizzo, Lindsay M Ross, Aidan G Ryan, Michelle A Niland, Laurence A Clarfeld, Stewart C Alexander, et al. Identifying connectional silence in palliative care consultations: a tandem machine-learning and human coding method. *Journal of Palliative Medicine*, 21(12):1755–1760, 2018.

[4] Lynn Bassett, Amanda F Bingley, and Sarah G Brearley. Silence as an element of care: A meta-ethnographic review of professional caregivers' experience in clinical and pastoral settings. *Palliative Medicine*, 32(1):185–194, 2018.

[5] Josef Bartels, Rachel Rodenbach, Katherine Ciesinski, Robert Gramling, Kevin Fiscella, and Ronald Epstein. Eloquent silences: A musical and lexical analysis of conversation between oncologists and their patients. *Patient Education and Counseling*, 99(10):1584–1594, 2016.

[6] Stephen A Buetow. Something in nothing: Negative space in the clinician-patient relationship. *The Annals of Family Medicine*, 7(1):80–83, 2009.

[7] Viktoria Manukyan, Brigitte N Durieux, Cailin J Gramling, Laurence A Clarfeld, Donna M Rizzo, Margaret J Eppstein, and Robert Gramling. Automated detection of conversational pauses from audio recordings of serious illness conversations in natural hospital settings. *Journal of Palliative Medicine*, 21(12):1724–1728, 2018.

[8] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg.

Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708, 2012.

[9] Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*, 2020.

[10] John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*, 2000.

[11] Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. Fisher english training speech part 1 transcripts. *Philadelphia: Linguistic Data Consortium*, 2004.

[12] Jan Svartvik. *The London–Lund corpus of spoken English: Description and research*, volume 82. 1990.

[13] Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University*, 2007.

# Comprehensive Bibliography


- Martin Aigner. *Combinatorial search*. John Wiley & Sons, Inc., 1988.
- Steven C Amstrup, Trent L McDonald, and Bryan FJ Manly. *Handbook of Capture-Recapture Analysis*. Princeton University Press, 2010.
- Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- Jordi Bascompte, Pedro Jordano, Carlos J Melián, and Jens M Olesen. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387, 2003.
- Ron Begleiter, Ran El-Yaniv, and Golan Yona. On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421, 2004.
- Michael GH Bell and Yasunori Iida. *Transportation Network Analysis*. 1997.
- Andrey Bernstein, Daniel Bienstock, David Hay, Meric Uzunoglu, and Gil Zussman. Power grid vulnerability to geographically correlated failures - analysis and control implications. In *INFOCOM, 2014 Proceedings IEEE*, pages 2634–2642. IEEE, 2014.
- Adam B Birchfield, Ti Xu, Kathleen M Gegner, Komal S Shetye, and Thomas J Overbye. Grid structural characteristics as validation criteria for synthetic networks. *IEEE Transactions on power systems*, 32(4):3258–3265, 2017.
- Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.
- Alexis A Borisy, Peter J Elliott, Nicole W Hurst, Margaret S Lee, Joseph Lehár, E Roydon Price, George Serbedzija, Grant R Zimmermann, Michael A Foley, Brent R Stockwell, et al. Systematic discovery of multicomponent therapeutics. volume 100, pages 7977–7982. National Acad Sciences, 2003.
- Ulrik Brandes, Garry Robins, Ann McCranie, and Stanley Wasserman. What is network science? *Network Science*, 1(1):1–15, 2013.
- Kenneth P Burnham and W Scott Overton. Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5):927–936, 1979.
- Kenneth L Calvert, Matthew B Doar, and Ellen W Zegura. Modeling internet topology. *IEEE Communications Magazine*, 35(6):160–163, 1997.

- Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.
- Anne Chao and Chun-Huo Chiu. Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, pages 1–26, 2014.
- Anne Chao and Lou Jost. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547, 2012.
- Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417):210–217, 1992.
- Jie Chen, James S Thorp, and Ian Dobson. Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model. *International Journal of Electrical Power & Energy Systems*, 27(4):318–326, 2005.
- Qiming Chen, Chuanwen Jiang, Wenzheng Qiu, and James D McCalley. Probability models for estimating the probabilities of cascading outages in high-voltage transmission network. *IEEE Transactions on Power Systems*, 21(3):1423, 2006.
- Ting Chen. A revised algorithm for searching for all defective edges in a graph. *Discrete Applied Mathematics*, 159(18):2266–2268, 2011.
- Ting Chen and Frank K Hwang. A competitive algorithm in searching for many edges in a hypergraph. *Discrete applied mathematics*, 155(4):566–571, 2007.
- Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.
- Chun-Huo Chiu, Yi-Ting Wang, Bruno A Walther, and Anne Chao. An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, 70(3):671–682, 2014.
- Wongyu Cho, Seong-Whan Lee, and Jin H Kim. Modeling and recognition of cursive words with hidden markov models. *Pattern Recognition*, 28(12):1941–1953, 1995.
- Monojit Choudhury and Animesh Mukherjee. The structure and dynamics of linguistic networks. In *Dynamics on and of Complex Networks*, pages 145–166. Springer, 2009.
- Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. Fisher english training speech part 1 transcripts. *Philadelphia: Linguistic Data Consortium*, 2004.

- Laurence A Clarfeld and Margaret J Eppstein. Group-testing on hypergraphs with variable-cost tests: A power systems case study. *arXiv preprint arXiv:1909.04513*, 2019.
- Laurence A Clarfeld, Margaret J Eppstein, Paul DH Hines, and Eric M Hernandez. Assessing risk from cascading blackouts given correlated component failures. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7. IEEE, 2018.
- Laurence A Clarfeld, Paul Hines, Eric Hernandez, and Maggie Eppstein. Risk of cascading blackouts given correlated component outages. *IEEE Transactions on Network Science and Engineering*, 2019.
- Charles J Colbourn. Reliability issues in telecommunications network planning. In *Telecommunications Network Planning*, pages 135–146. Springer, 1999.
- Jin Cong and Haitao Liu. Approaching human language with complex networks. *Physics of Life Reviews*, 11(4):598–618, 2014.
- Rong-Gang Cong and Mark Brady. The interdependence between rainfall and temperature: copula analyses. *The Scientific World Journal*, 2012, 2012.
- George Constable and Bob Somerville. *A Century of Innovation: Twenty Engineering Achievements that Transformed our Lives*. Joseph Henry Press, 2003.
- 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.
- National Research Council et al. *Network Science*. National Academies Press, 2006.
- Marco Cristani, Anna Pesarin, Carlo Drioli, Alessandro Tavano, Alessandro Perina, and Vittorio Murino. Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8):1785–1800, 2011.
- Alfredo Cuzzocrea, Alexis Papadimitriou, Dimitrios Katsaros, and Yannis Manolopoulos. Edge betweenness centrality: A novel algorithm for qos-based topology control over wireless sensor networks. *Journal of Network and Computer Applications*, 35(4):1210–1217, 2012.
- Peter Damaschke. A tight upper bound for group testing in graphs. *Discrete Applied Mathematics*, 48(2):101–109, 1994.
- Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708, 2012.

- Martijn de Jong, Georgios Papaefthymiou, and Peter Palensky. A framework for incorporation of infeed uncertainty in power system risk-based security assessment. *IEEE Transactions on Power Systems*, 33(1):613–621, 2018.

- 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.

- Ian Dobson, NichelleLe K Carrington, Kai Zhou, Zhaoyu Wang, Benjamin A Carreras, and José M Reynolds-Barredo. Exploring cascading outages and weather via processing historic data. In *Hawaii International Conference on System Sciences (HICSS-51)*, Jan 2017.

- Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

- Dingzhu Du and Frank Hwang. *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

- Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial Group Testing and its Applications*, volume 12. World Scientific, 2000.

- Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

- Fabrizio Durante and Carlo Sempi. *Principles of Copula Theory*. CRC press, 2015.

- Valdo Durrleman, Ashkan Nikeghbali, and Thierry Roncalli. Which copula is the right one? *Available at SSRN 1032545*, 2000.

- Margaret J Eppstein and Paul DH Hines. A "random chemistry" algorithm for identifying collections of multiple contingencies that initiate cascading failure. *IEEE Transactions on Power Systems*, 27(3):1698–1705, 2012.

- Margaret J Eppstein, Joshua L Payne, Bill C White, and Jason H Moore. Genomic mining for complex disease traits with 'random chemistry'. *Genetic Programming and Evolvable Machines*, 8(4):395–411, 2007.

- Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Petropolitanae*, pages 128–140, 1741.

- Betty Ferrell, Stephen R Connor, Anne Cordes, Constance M Dahlin, Perry G Fine, Nancy Hutton, Mark Leenay, Judy Lentz, Judi Lund Person, Diane E Meier, et al. The national agenda for quality palliative care: the national consensus project and the national quality forum. *Journal of pain and symptom management*, 33(6): 737–744, 2007.

- Gernot A Fink. *Markov Models for Pattern Recognition: From Theory to Applications*. Springer Science & Business Media, 2014.

- Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

- Christian Gollier and Olivier Gossner. Group testing against covid-19. *Covid Economics*, 2, 2020.

- Simeon Hagspiel, Antonis Papaemannouil, Matthias Schmid, and Göran Andersson. Copula-based modeling of stochastic wind power in europe and implications for the swiss power grid. *Applied energy*, 96:33–44, 2012.

- Paul Haidet and Debora A Paterniti. Building a history rather than taking one: A perspective on information sharing during the medical interview. *Archives of internal medicine*, 163(10):1134–1140, 2003.

- James F Heltshe and Nancy E Forrester. Estimating species richness using the jackknife procedure. *Biometrics*, pages 1–11, 1983.

- Patrick Heneka and Bodo Ruck. A damage model for the assessment of storm damage to buildings. *Engineering Structures*, 30(12):3603–3609, 2008.

- Paul Hines and Pooya Rezaei. *Smart Grid Handbook*, chapter Cascading Failures in Power Systems. John Wiley & Sons, 2016.

- Michael Kai Hourfar, Anna Themann, Markus Eickmann, Pilaipan Puthavathana, Thomas Laue, Erhard Seifried, and Michael Schmidt. Blood screening for influenza. *Emerging Infectious Diseases*, 13(7):1081, 2007.

- Jacqueline M Hughes-Oliver. Pooling experiments for blood screening and drug discovery. In *Screening*, pages 48–68. Springer, 2006.

- Frank K Hwang. A competitive algorithm to find all defective edges in a graph. *Discrete applied mathematics*, 148(3):273–277, 2005.

- Hiroko Itakura. Describing conversational dominance. *Journal of Pragmatics*, 33 (12):1859–1880, 2001.

- Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle, and Tomasz Rychlik. *Copula Theory and its Applications*, volume 198. Springer, 2010.
- Nirmal Jayaram and Jack W Baker. Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics*, 38(15): 1687–1708, 2009.
- Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- Kai Jiang and Chanan Singh. New models and concepts for power system reliability evaluation including protection system failures. *IEEE Transactions on Power Systems*, 26(4):1845–1855, 2011.
- Petra Johann. A group testing problem for graphs with several defective edges. *Discrete Applied Mathematics*, 117(1-3):99–108, 2002.
- Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, and Chris Sander. Predicting protein structure using hidden markov models. *Proteins: Structure, Function, and Bioinformatics*, 29(S1): 134–139, 1997.
- Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):121–125, 1999.
- Richard W Katz. On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249, 1981.
- Stuart Kauffman. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford university press, 1996.
- HB Kekre, CL Saxena, and HM Srivastava. A two-state markov model of speech in conversation and its application to computer communication systems. *Computers & Electrical Engineering*, 4(2):133–141, 1977.
- Sun Bean Kim, Hye Won Kim, Hyon-Suk Kim, Hea Won Ann, Jae Kyoung Kim, Heun Choi, Min Hyung Kim, Je Eun Song, Jin Young Ahn, Nam Su Ku, et al. Pooled nucleic acid testing to identify antiretroviral treatment failure during hiv infection in seoul, south korea. *Scandinavian Journal of Infectious Diseases*, 46(2): 136–140, 2014.

- Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.
- Klaus Köck, Herwig Renner, and Josef Stadler. Probabilistic cascading event risk assessment. In *Power Systems Computation Conference (PSCC), 2014*, pages 1–7. IEEE, 2014.
- Ann E Krause, Kenneth A Frank, Doran M Mason, Robert E Ulanowicz, and William W Taylor. Compartments revealed in food-web structure. *Nature*, 426 (6964):282–285, 2003.
- Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001.
- Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.
- Jérôme Lapuyade-Lahorgue, Jing-Hao Xue, and Su Ruan. Segmenting multi-source images using hidden markov fields with copula-based multivariate statistical distributions. *IEEE Transactions on Image Processing*, 26(7):3187–3195, 2017.
- Patrick Laux, S Vogl, W Qiu, Hans Richard Knoche, and Harald Kunstmann. Copula-based statistical refinement of precipitation in rcm simulations over complex terrain. *Hydrology and Earth System Sciences*, 15:2401–2419.
- Wenyuan Li. Incorporating aging failures in power system reliability evaluation. *IEEE Transactions on Power systems*, 17(3):918–923, 2002.
- Wenyuan Li. *Risk assessment of power systems: models, methods, and applications*. John Wiley & Sons, 2014.
- Smriti Mallapaty. The mathematical strategy that could transform coronavirus testing. *Nature*, 583(7817):504–505, 2020.
- John D Musa and Kazuhira Okumoto. A logarithmic poisson execution time model for software reliability measurement. In *Proceedings of the 7th International Conference on Software Engineering*, pages 230–238. Citeseer, 1984.
- Roger B Nelsen. *An introduction to copulas*. Springer, New York, 2010. ISBN 978-1-4419-2109-3 978-0-387-28678-5.

- D.E. Newman, B.A. Carreras, V.E. Lynch, and I. Dobson. Exploring complex systems aspects of blackout risk and mitigation. *IEEE Transactions on Reliability*, 60(1):134 –143, 2011.
- Institute of Medicine (US). Committee on Approaching Death: Addressing Key End-of Life Issues. *Dying in America: Improving quality and honoring individual preferences near the end of life*. National Academies Press, 2015.
- RB O'hara. Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, pages 375–386, 2005.
- Jens M Olesen, Jordi Bascompte, Yoko L Dupont, and Pedro Jordano. The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891–19896, 2007.
- Arno Onken, Steffen Grünewälder, Matthias HJ Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS computational biology*, 5(11):e1000577, 2009.
- Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 191–198. ACM, 2005.
- Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.
- George Papaefthymiou and Dorota Kurowicka. Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE Transactions on Power Systems*, 24(1):40–49, 2009.
- George Papaefthymiou and Pierre Pinson. Modeling of spatial dependence in wind power forecast uncertainty. In *Probabilistic Methods Applied to Power Systems, 2008. PMAPS'08. Proceedings of the 10th International Conference on*, pages 1–9. IEEE, 2008.
- Milorad Papic, Sudhir Agarwal, Ron N Allan, Roy Billinton, Chris J Dent, Svetlana Ekisheva, Daniel Gent, Kai Jiang, Wenyuan Li, Joydeep Mitra, et al. Research on Common-Mode and Dependent (CMD) outage events in power systems: A review. *IEEE Transactions on Power Systems*, 32(2):1528–1536, 2017.

- John R Papp, Julius Schachter, Charlotte A Gaydos, and Barbara Van Der Pol. Recommendations for the laboratory-based detection of chlamydia trachomatis and neisseria gonorrhoeae - 2014. *MMWR. Recommendations and Reports: Morbidity and mortality weekly report. Recommendations and Reports/Centers for Disease Control*, 63:1, 2014.

- Kevin C Parker. Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965, 1988.

- Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University*, 2007.

- Jennifer Pohle, Roland Langrock, Floris M van Beest, and Niels Martin Schmidt. Selecting the number of states in hidden markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):270–293, 2017.

- Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

- Katja S Remlinger, Jacqueline M Hughes-Oliver, S Stanley Young, and Raymond L Lam. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, 48(1):133–143, 2006.

- Pooya Rezaei, Paul DH Hines, and Margaret Eppstein. Estimating cascading failure risk: Comparing monte carlo sampling and random chemistry. In *PES General Meeting | Conference & Exposition, 2014 IEEE*, pages 1–5. IEEE, 2014.

- Pooya Rezaei, Margaret J Eppstein, and Paul DH Hines. Rapid assessment, visualization, and mitigation of cascading failure risk in power systems. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 2748–2758. IEEE, 2015.

- Pooya Rezaei, Paul DH Hines, and Margaret J Eppstein. Estimating cascading failure risk with random chemistry. *IEEE Transactions on Power Systems*, 30(5): 2726–2735, 2015.

- Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier, 1978.

- Abdullahi M Salman. Age-dependent fragility and life-cycle cost analysis of timber and steel distribution poles subjected to hurricanes. Master's thesis, Michigan Technological University, 2014.
- Anke Scherb, Luca Garrè, and Daniel Straub. Reliability and component importance in networks subject to spatially distributed hazards followed by cascading failures. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 3(2):021007, 2017.
- Christian Schoelzel and Petra Friederichs. Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15(5):761–772, 2008.
- Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 2, pages II–1. IEEE, 2003.
- Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- Bryan Severyn, Robert A Liehr, Alex Wolicki, Kevin H Nguyen, Edward M Hudak, Marc Ferrer, Jeremy S Caldwell, Jeffrey D Hermes, Jing Li, and Matthew Tudor. Parsimonious discovery of synergistic drug combinations. *ACS Chemical Biology*, 6 (12):1391–1398, 2011.
- Rebecca Smith. Us risks national blackout from small-scale attack. *Wall Street Journal*, 12, 2014.
- Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-based Recognition*, pages 227–243. Springer, 1997.
- Jan Svartvik. *The London-Lund corpus of spoken English: Description and research*. Number 82. Lund University Press, 1990.
- Scott M Thede and Mary Harper. A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 175–182, 1999.
- EG Tiedemann. Channel access protocols for half duplex satellite terminals. In *IEEE Military Communications Conference,'Bridging the Gap. Interoperability, Survivability, Security'*, pages 463–469. IEEE, 1989.

- Howell Tong. Determination of the order of a markov chain by akaike's information criterion. *Journal of Applied Probability*, 12(3):488–497, 1975.

- Eberhard Triesch. A group testing problem for hypergraphs of bounded rank. *Discrete Applied Mathematics*, 66(2):185–188, 1996.

- Marianna Vaiman, Keith Bell, Yousu Chen, Badrul Chowdhury, Ian Dobson, Paul Hines, Milorad Papic, Stephen Miller, and Pei Zhang. Risk assessment of cascading outages: Methodologies and challenges. *IEEE Transactions on Power Systems*, 27 (2):631, 2012.

- Alexei Vázquez, Romualdo Pastor-Satorras, and Alessandro Vespignani. Large-scale topological and dynamical properties of the internet. *Physical Review E*, 65(6): 066130, 2002.

- Stanley Wasserman and Katherine Faust. Social network analysis in the social and behavioral sciences. *Social Network Analysis: Methods and Applications*, 1994:1–27, 1994.

- HE Willoughby, RWR Darling, and ME Rahn. Parametric representation of the primary hurricane vortex. part ii: A new family of sectionally continuous profiles. *Monthly Weather Review*, 134(4):1102–1120, 2006.

- Andreas Wulvik, Matilde Bisballe Jensen, and Martin Steinert. Temporal static visualisation of transcripts for pre-analysis of video material: Identifying modes of information sharing. In *Analysing Design Thinking: Studies of Cross-Cultural Co-Creation*. CRC Press, 2017.

- Joshua Wurman and Curtis R Alexander. The 30 may 1998 spencer, south dakota, storm. part ii: Comparison of observed damage and radar-derived winds in the tornadoes. *Monthly Weather Review*, 133(1):97–119, 2005.

- Lan Zhang and Vijay P Singh. *Copulas and Their Applications in Water Resources Engineering*. Cambridge University Press, 2019.

- Yingbo Zhou, Utkarsh Porwal, Ce Zhang, Hung Q Ngo, XuanLong Nguyen, Christopher Ré, and Venu Govindaraju. Parallel feature selection inspired by group testing. In *Advances in Neural Information Processing Systems*, pages 3554–3562, 2014.

- Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, Robert John Thomas, et al. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, 26(1):12–19, 2011.