UNIVERSITY OF CALIFORNIA
Santa Barbara

Information Propagation on Social Networks

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Mechanical Engineering

by

Michael J. Busch Jr.

Committee in Charge:

Professor Jeff Moehlis, Chair

Professor Francesco Bullo

Professor Igor Mezić

Professor João Hespanha

December 2014

The Dissertation of
Michael J. Busch Jr. is approved:

Professor Francesco Bullo

Professor Igor Mezić

Professor João Hespanha

Professor Jeff Moehlis, Committee Chairperson

October 2014

Information Propagation on Social Networks

by

Michael J. Busch Jr.

To anyone willing to read this manuscript.

Acknowledgements

First and foremost I would like to thank my advisor, Jeff Moehlis, for being an awesome mentor and academic tour guide over the past few years. In addition, I am grateful for the guidance and mentorship of Ambuj Singh, and especially Petko Bogdanov, who contributed much of the original material in Section 3.1, including Figures 3.1, 3.2, 3.3, 3.4, and their accompanying discussion.

This journey of obtaining a PhD has been arduous and lonesome at times, and I am forever grateful to have had some amazing friends and colleages at my side. Specifically, I would like to thank the past and present members of the DCR Lab including Michael Nip, Lina Kim, Ryan Mohr, Marco Budisic, Patrick Shepherd, Ali Nabi, Margot Kimura, Per Danzl, Louis Van Blarigan, Mitchel Craun, and the Pacific Ocean. And to everyone else who I do not have space to specifically mention (you know who you are), I thank you too.

Most importantly, I need to thank Mom and Dad for their enduring support and patience since July 27th, 1986. Lastly, I would like to acknowledge my highschool math teacher Mr. Bradlee, my undergraduate Professors at RPI, Henry Scarton and Theo Borca-Tasciuc, and my former supervisor at ViaSat, Derek Johnson. I wouldn't be here if it weren't for their support and recommendations over the years.

Funding

Much of this work was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein. The content of the information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Curriculum Vitæ

Michael J. Busch Jr.

Education

2008–2014 Ph.D. Candidate in Mechanical Engineering, University of California, Santa Barbara, CA.

2004–2008 Bachelor of Science in Mechanical Engineering, Minor in Electrical Engineering, Rensselaer Polytechnic Institute, Troy, NY.

Academic Work Experience

2009–2014 **Graduate Student Researcher**, UC Santa Barbara, Research group of Dr. Jeff Moehlis.

2014 **Research Mentor**, UC Santa Barbara Summer Sessions Research Mentorship Program.

2014 **Research Mentor**, UC Santa Barbara IGERT.

2013 **Course Instructor**, UC Santa Barbara, ENGR 3 (Introduction to Computer Programming).

2012 **Teaching Assistant**, UC Santa Barbara, ENGR 3 (Introduction to Computer Programming).

2011 **Research Mentor**, UC Santa Barbara Institute for Collaborative Biotechnologies SABRE Program.

2009, 2012 **Teaching Assistant**, UC Santa Barbara, ME 6 (Circuits).

2008 **Teaching Assistant**, UC Santa Barbara, ME 104 (Mechatronics).

2007–2008 **Undergraduate Student Researcher**, Rensselaer Polytechnic Institute, Research group of Dr. Henry Scarton.

2007 **Undergraduate Student Researcher**, Rensselaer Polytechnic Institute, Research group of Dr. Theodorian Borca-Tasciuc.

Industry Work Experience

2012 **Market Researcher**, MyCarna Inc., UC Santa Barbara Technology Management Program.

2007, 2008 **Engineering Intern**, ViaSat Inc., Germantown, MD.

2006 **Micro-fabrication Technician**, Holographix LLC., Hudson, MA.

Journal Publications (peer reviewed)

- P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski. Modeling individual topic-specific behavior and influence backbone networks in social media. Social Network Analysis and Mining, 4:204, 2014.

- A. Kolpas, M. Busch, H. Li, I.D. Couzin, L. Petzold, J. Moehlis. How the Spatial Position of Individuals Affects Their Influence on Swarms: A Numerical Comparison of Two Popular Swarm Dynamics Models. PLoS ONE 8(3): e58525, 2013.

- M. Busch and J. Moehlis. On the homogeneous assumption and the logistic behavior of information propagation. Physical Review E 85:026102, 2012.

- M. Busch and J. Moehlis. Analysis of a class of symmetric equilibrium configurations for a territorial model. Numerical Mathematics: Theory, Methods, and Applications 3, 143-161, 2010.

Conference Proceedings (peer reviewed)

- M. Busch and J. Moehlis, A nonparametric adaptive nonlinear statistical filter. Proceedings of the 53rd IEEE Conference on Decision and Control (CDC), Los Angeles, CA. (Accepted).

- P. Bogdanov, M. Busch, J. Moehlis, A. Singh and B. Szymanski. The social media genome: modeling individual topic-specific behavior in social media. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Niagara Falls, Canada.

Selected Presentations

- "The social media genome: modeling individual topic-specific behavior in social media." IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 25-28, 2013, Niagara Falls, Canada.

- "Information Propagation Models and Social Networks." SIAM Conference on Applications of Dynamical Systems, May 22-26, 2011, Snowbird, UT.

- M. Busch and J. Moehlis, Modeling rumor propagation through heterogeneous networks. Proceedings of 2010 Network Science Workshop, West Point, New York.

Selected Posters

- "Agent-Based Description of the Homogeneous Assumption," 2011 Complexity Conference, Northwestern Institute on Complex Systems (NICO), March 6-7, 2011, Evanston, Illinois.

- "What's the buzz? A coarse description of information propagation." Institute for Collaborative Biotechnologies (ICB) Army-Industry Collaboration Conference, February 8-9, 2011, Santa Barbara, CA.

Awards & Honors

- Honorary IGERT Fellow (2013)
- Excellence Fellowship, UCSB Dept. of Mech. Eng. (2012)
- Team semifinalist, UCSB New Venture Competition (2012)
- Magna Cum Laude, Bachelor of Science (2008)
- Tau Beta Pi, all engineering honor society (2007)
- Pi Tau Sigma, mechanical engineering honor society (2006)
- BAE Systems LITEC Challenge: 2nd Place (2006)
- Academic Citation, Differential Equations (2005)
- Phi Mu Delta (2004), Chapter President (2006)
- Rensselaer Medal (2003)

Abstract

Information Propagation on Social Networks

Michael J. Busch Jr.

Many models of disease and rumor spreading phenomena average the behavior of individuals in a population in order to obtain a coarse description of expected system behavior. For these types of models, we determine how close the coarse population-level approximation is to its corresponding agent-based system and discuss the accuracy of the population-level approximation. We apply these theoretical results to real social network data to see how well they describe the contagious nature of social phenomena. Specifically, we consider hashtag adoption data collected from the Twitter social network. To assimilate the Twitter data to a simple contagion model, we developed and implemented statistical learning methods to construct an adaptive state estimator for systems described by nonlinear stochastic differential equations.

We found that the static network structure alone is not sufficient for explaining hashtag adoption among users in the Twitter social network, and our result suggest that a user-centric model would be more appropriate for this task. We propose a model for individual social media users, termed a *genotype*, which is a *per-topic* summary of a user's interest, activity and susceptibility to adopt new information. We show that the genotype framework is capable of accurately quantifying the adoption behavior of individual users with respect to hashtag topics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We often speak about the latest YouTube video or cat meme as going *viral*, but it is often overlooked how similar the sharing of social content, whether in real life or on the internet, is to the sharing of an infectious disease among individuals in a population. For both infectious diseases and social phenomena, each individual is a host who, in general, carries a parcel of information and passes it on to others. Although the parcels of information and transmission mechanisms may be different, transmission events in both cases are constrained to occur only between those people who come in contact with those who are infected/informed.

Fortunately with the recent emergence of massive on-line social networks, such as Facebook and Twitter, there exist not only social platforms for information to be shared across the globe, but also a well defined network structure on which this social information spreads. Therefore, this manuscript will investigate the significance of the social network structure on the sharing of social information in a population, and uses real data from

the Twitter social network to evaluate the efficacy of agent-based network models in predicting the adoption of hashtags among the Twitter population.

Contagion models from epidemiology are introduced in Chapter 2 and applied to agent-based contact networks. These contact networks can be studied using the tools from the field of algebraic graph theory, and population-level comparisons can be made with prior results in the literature, which rely on statistical mechanics. The main results of Chapter 2 include a discussion of the specific case when the network results and the statistical mechanics results are in exact agreement, and show how deviations from the statistical mechanics results are attributed to the network structure.

The Twitter social network and real Twitter data are introduced in Chapter 3. For simplicity we will focus our attention on population-wide user adoption behavior of *hashtags*, which are user generated tokens that begin with the # symbol. Since the Twitter social network is so large and contains only a subset of each individual's full social network, we also introduce the notion of *backbone* network structures in order to accommodate these shortcomings of the available Twitter network structure.

In Chapter 4 and Chapter 5, we study the ability of the fine-grained network models of Chapter 2 to explain the observed population-level hashtag adoption behavior in the Twitter network. Specifically, in Chapter 4 we develop a method for estimating ensemble statistics of a stochastic model from a single realization of time-series data. This method is applied to the Twitter hashtag data in Chapter 5, and comparisons are made between the Twitter data and the network models. The results of Chapter 5 point to some of

the shortcomings of the agent-based network models, and motivates the need for a more descriptive social network user model.

A genetically inspired user model, called a *social genotype* is introduced and described in Chapter 6 as a more descriptive social network user model. We demonstrate how this genotype can be constructed, and show how it is invariant for each user. We show that when used in conjunction with the Twitter backbone structures of Chapter 3, the Twitter genotype of each user can accurately predict hashtag adoption at the topic-level.

Prior Publications

Much of the work contained in this manuscript is also contained in our earlier peer-reviewed publications. Specifically, the content of Chapter 2 also appears in [2], while the content of Chapter 3 and Chapter 6 also appears in [3] and [4]. Chapter 4 contains material from [5], and the content of Chapter 5 is original to this manuscript. We have obtained written copyright permission from the original publishers, and we have reproduced all figures and text in accordance with policy of both the University of California and the original publishers.

Chapter 2

Population and Network Models

Many different phenomena can generally be described as the exchange of information between members of a population, such as the spread of rumors [6–9], ideas [10], computer worms and viruses over the internet [11, 12], and most notably the spread of infectious diseases [13–25]. These types of phenomena can be modelled at the population level, the agent level, or somewhere in between. Moreover, the popular work on small-world [1] and scale-free [26] networks have uncovered important structural details of typical populations in these systems. Recent discoveries in the discipline of network science have motivated a second look at how structural details of the population network influence the rate and depth at which information spreads, and how agent-level interactions affect population-level interactions [18–25]. Here, we continue along these lines and draw attention to novel ways of quantitatively analyzing how network structure affects the spread of information through a population.

Early descriptions of disease and rumor propagation approximate population level behavior by low dimensional ordinary differential equations (ODEs) [6,13–15,17]. However,

it is difficult to include the network topology of the propagation medium in these low dimensional models. A related issue is how local variations in behavior and connectivity should be averaged, which is known as the problem of heterogeneity [17, 21, 27]. The low dimensional models implicitly assume that every node in the network is connected to every other node in the network equally [28]. Initial attempts at addressing this problem include proportional mixing techniques, which focused on subdividing the population into smaller homogeneous populations [16]. Recently, the inclusion of network topology has been addressed using heterogeneous mean-field approaches that coarse-grain the set of nodes into various degree classes that have similar dynamical properties [18,22–24,29,30]. These methods rely on statistical properties of the network rather than the global structure of the network as a whole, and become computationally costly as the system detail is refined to the agent level.

In this chapter we shall analyze the spread of information between individuals for simple contagion scenarios [31], and develop an agent-based propagation model that is similar to the probabilistic discrete-time Markov chain studied in [32]. This agent-based theoretical framework scales to subpopulations of any size [13] and has been shown to generalize heterogeneous mean field approaches [32], where the subpopulations are typically groups of nodes that have the same out-degree. The advantage to using an agent-based theoretical framework is that it begins with the exact structure of the network and determines the dynamics rather than incorporating the detailed network properties into an already formulated coarse dynamic model, thus allowing one to accurately probe

network effects at any scale. Hence, we seek to show that our agent-based contact model is consistent with the well-known low dimensional mean-field logistic model, and discuss the implications of using a low dimensional logistic model in place of the agent-based contact model. Because it is popularly taken for granted that low dimensional models are not accurate representations of agent-based systems [33], our approach will attempt to rigorously quantify the differences between the two representations and uncover the network topologies where the low dimensional and agent-based models may possibly agree.

Our findings rely on the implementation of algebraic graph theory, which has been extensively applied to the analysis of static network structure [34], particularly since adjacency matrices uniquely define a graph [34] and are computationally efficient mathematical structures [35]. From our agent-based approach to the study of information propagation, we argue that new physical insight is gained by applying the tools of algebraic graph theory to the study of the dynamics on a network. Not only will adjacency matrices allow us to rigorously attribute the size of fluctuations about the population-level solutions to finite size effects, but they will simultaneously tell us what network structures are necessary for our assumptions to hold. Furthermore, since the the mean-field population model implicitly asserts a particular graph structure, i.e. a completely connected graph, we will conclude our discussion by exploring how one can use a mean-field solution for an agent-based system, which may have an arbitrary graph structure, to assert parameter values of a corresponding coarse population model. We propose a novel

way in which one can use these inferred parameter values as simple metrics for comparing the graph-dependent propagation dynamics for a set of initially informed nodes.

To communicate these ideas, this chapter is organized as follows: in Section 2.1, a coarse-grain model based on the scalar logistic equation is introduced. We present the agent-based contact model in Section 2.2; this describes the probability of interaction between nodes on the graph in terms of the exact network topology. Section 2.3 describes how the models developed in Sections 2.1 and 2.2 intersect, which leads to a general result for doubly stochastic networks and networks with interaction symmetry that is presented in Section 2.4. The logistic behavior of example heterogeneous networks is discussed in Section 2.5, and a summary of our findings is given in Section 2.6.

2.1 Logistic Population Model

For a given population of N agents, suppose each agent can only be a member of one of two sets: the set of S-class individuals (susceptibles) or the set of I-class individuals (informeds). Agents can only go from being susceptible to being informed, and once informed remain informed for all future time. If one were to only consider the transfer events of information propagation, then the SI model is arguably the simplest population-level model that corresponds with information communication mechanisms between members. Including more complicated behavior by allowing members to leave the I-class set, by either reentering the S-class or entering an R-class sub-population (removed/forget), is unnecessary because these are behaviors that do not typically depend

on network structure; it is not necessarily true that one's neighbors' forgetfulness or recovery will directly cause one to also forget or recover. For these reasons, we will simplify our discussion by only considering SI dynamics.

The classic SI model [13, 15, 17] is expected to only be accurate for systems that exhibit well-mixed behavior because it assumes that (i) the population size is fixed, and (ii) the members within a set are indistinguishable from every other member in that set. In discrete time, the rate of infection is proportional to the number of susceptibles and informeds at the previous time step, as well as a transmission rate β, which we shall generally treat as a time varying expression, β_t [36]. By letting I_t be the proportion of informed individuals in a population, the SI dynamics are described by the discrete scalar difference equation:

$$I_{t+h} = I_t + h\beta_t I_t \left(1 - I_t\right), \tag{2.1}$$

and the *logistic* function that solves this equation.

This model has the additional assumption that (iii) the time step h is small enough such that only one informed agent contacts all of his neighbors during that time step. As noted in [13], solutions to (2.1) are bounded on the unit interval for each initial value on the unit interval only if the coefficient of the $I_t\left(1 - I_t\right)$ term, say $\alpha(t)$, is positive and satisfies the condition

$$\sup_t \alpha(t) \leq 1. \tag{2.2}$$

For (2.1), $\alpha(t) = h\beta_t$, and this implies that the step size of (2.1) must satisfy $h \leq 1/\sup_t \beta_t$.

2.2 Agent-Based Model

As an alternative, let us now consider a system of *discrete agents* that are able to share information with each other. The communication pathways between agents can be mapped as a graph $G(V, E)$ of N total agents [28, 37], where each uniquely indexed node $i \in \{1, \ldots, N\} \subset V$ of the graph represents a distinct agent, and a directed edge $(i, j) \in E$ connecting two nodes i and j indicates that it is possible for agent i to transmit information to agent j. For example, in epidemiology an agent is a unique person and a parcel of information may be an infectious disease [38], and in the blogosphere an agent is a unique web user while a parcel of information may be a specific rumor about a politician or celebrity [39].

We are interested in the probability that agent i is in possession of a specific parcel of information at discrete time t, which we denote $p_t^{(i)}$. Furthermore, agent i is assumed to communicate with a neighbor, say agent j with $j \neq i$, in such a way that there is a nonzero probability a_{ij} that agent j successfully communicates a parcel of information to agent i in a time period of h. In general, each a_{ij} is a time-varying expression since the graph topology of a given social network is subject to change over long enough periods of time. To keep our discussion simple, we adopt the common assumption that the information of interest spreads through the network faster than significant changes to the network are able to emerge.

Motivated by the mechanism of social media platforms such as Twitter and the Facebook newsfeed, the magnitude of each a_{ij} for a given i is the probability that, in one

time step, agent i will be contacted by one of his neighbors j. In this sense, the informed agents broadcast the information to their neighbors. Thus, a_{ij} is said to be an element of the weighted adjacency matrix A that uniquely defines the structure of the graph G [37], and each $a_{ij} \in \left\{ [0,1] : \forall\, i \in \{1,\ldots,N\},\ \sum_{j=1}^{N} a_{ij} = 1 \right\}$. Any matrix that has this property is said to be row stochastic.

Recent evidence suggests the probability that an information transfer event occurs is also dependent on the nature of the information itself [39]. For instance, two agents of a network may be in communication, but not necessarily sharing the type of information that one would like to be tracking. Thus, the average probability that the desired information is being transmitted during a given period of time h is given by $h\beta_t$, where β_t is similar to the transmission rate defined for scalar logistic models. We remark that the step size requirement (2.2) ensures the term $h\beta_t$ abides the probability axiom $h\beta_t \in [0,1]$.

It shall be assumed that once an individual is informed (infected), he does not forget (recover) or become silent (removed). Instead, we attribute any time varying effects of the propagation dynamics to the nature of the information itself through the β_t expression. Given these conditions, the total probability of an arbitrary agent i becoming informed at a given time step $t+h$, denoted by $p_{t+h}^{(i)}$, is

$$p_{t+h}^{(i)} = p_t^{(i)} + \left(1 - p_t^{(i)}\right) \left(\sum_{j=1}^{N} h\beta_t a_{ij} p_t^{(j)} \right). \tag{2.3}$$

Hence, an agent is informed at time $t+h$ if he is already informed by time t, or the agent is not informed by time t and an informed neighbor successfully transmits the information. The fact that the probability at the next time step only depends on the

probability at the current time step indicates that the system of equations expressed by (2.3) has the Markov property, and is referred to as a Markov chain [40]. We remark that the diagonal elements a_{ii} of the weighted adjacency matrix are necessarily equal to zero. If this were not the case, then a contradiction would occur because then an uninformed agent would be able to spontaneously inform himself. In matrix notation, (2.3) becomes

$$p_{t+h} = p_t + h\beta_t \left(I - diag\left\{p_t\right\}\right) A p_t, \tag{2.4}$$

where p_t is a column vector whose indices correspond with the agent indices.

Epidemic models of this form have been shown by Monte Carlo simulation to generalize both contact processes and reactive processes [32]. A contact process is a dynamical process where each informed agent stochastically informs just one of his neighbors per time step, while a reactive process is a dynamical process where at least one informed agent stochastically informs all of his neighbors per time step. Hence, by construction, we consider the dynamics of a reactive process.

Since we consider the dynamics of a reactive process, we constrain each S-class individual to only interact with one of his neighbors at a time. This assumption is consistent with the mechanics of simple contagions, and applies to situations where information is broadcast, say a radio signal for example, and each susceptible individual can only "listen" to one broadcasting source at a time so that the communication events are mutually exclusive. In contrast to equation (2) of [32], where elements of the weighted adjacency matrix describe the probability of where a random walker on the network will go next,

the elements of the weighted adjacency matrix in our system describe the probability of where the random walker has come from. When comparing these two closely related frameworks, equation (2.3) of this chapter can be recovered from equation (1) of [32] by first swapping the index of the product, and then applying De Morgan's law to obtain a series representation.

2.3 Completely Connected Solution

Having introduced both a population-level model and an agent-based model independently, one can rigorously construct the population-level dynamics directly from the agent-level dynamics by asserting the "well-mixed" assumption that is implicit in the population-level dynamics [28] presented in Section 2.1. In terms of graph topology, we argue that well-mixedness of a population corresponds to a completely connected graph. A graph is said to be completely connected if every node on the graph shares an undirected edge with every other node on the graph [37], and a network of agents on a completely connected graph is often considered to be "well-mixed" if every agent communicates with every other agent equally [28]. A graph of this type is described as being "homogeneous" because the local graph topology for each node is indistinguishable from that of every other node. For a network of N agents, this implies that elements of the weighted adjacency matrix for a completely connected graph have the following

values

$$a_{ij} = \begin{cases} \frac{1}{N-1}, & i \neq j \\ \\ 0, & i = j \end{cases}.$$

We remark that a model in this framework, as stated, relies on the assumption that (i) the population does not change, and (ii) the edge weights do not change. The regularity of the adjacency matrix for the well-mixed case allows one to also find upper and lower bounding functions to the solution of (2.4). Since the solution to (2.3) is positive and monotonically increasing element-wise [13], taking the one norm is identical to summing over all of the elements:

$$\begin{aligned} |p_{t+h}|_1 &= \sum_{i=1}^{N} \left(p_t^{(i)} + h\beta_t \left(1 - p_t^{(i)}\right) \left(\sum_{j \neq i} \frac{1}{N-1} p_t^{(j)} \right) \right) \\ &= |p_t|_1 + h\beta_t |p_t|_1 \left(1 - \frac{|p_t|_1}{N-1}\right) + \frac{h\beta_t}{N-1} |p_t|_2^2 . \end{aligned} \tag{2.5}$$

We note that the maximum possible informed population - also known as carrying capacity [14] - of the model (2.5) is N, rather than $N-1$ as the resemblance of (2.5) to the discrete logistic equation might falsely suggest.

It is now possible to compare the graph based solution of equation (2.4) to that of the traditional Susceptible-Infected (SI) model for systems containing an arbitrary number of agents. If the cardinality of the susceptible and infected populations are random variables, then experimental evidence suggests that the scalar variables of the low dimensional models represent the expected values for the sizes of those sets [13].

For comparison, equation (2.5) can be normalized with respect to the total population to obtain

$$x_{t+h} = x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right) + \frac{h\beta_t}{N\,(N-1)} \left|p_t\right|_2^2, \tag{2.6}$$

where $x_t = \left|p_t\right|_1 / N$ has the usual interpretation of being the expected probability that an arbitrarily sampled agent is informed. Under this interpretation, one can think of (2.6) as a mean-field description of the population. Another interpretation of x_t is that it represents the proportion of informed individuals in a population.

In the thermodynamic limit where the size of the system, N, approaches infinity, one finds that

$$x_{t+h} = x_t + h\beta_t x_t \left(1 - x_t\right). \tag{2.7}$$

Hence, the dynamics of the SI model and the graph-based model are equivalent in the thermodynamic limit. Given a population of size N, the solutions to the difference equations (2.1) and (2.7) are equal when the initial concentration of (2.7) is taken to be the proportion of initially informed individuals of (2.1). For finite homogeneous populations, however, the solutions are closely bounded by the solutions to

$$x_{t+h} = x_t + h\beta_t \frac{N}{N-1} x_t \left(1 - x_t\right), \qquad \text{and} \qquad x_{t+h} = x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right),$$

when the following step-size condition is met:

$$h \leq \frac{N-1}{N+1} \frac{1}{\sup_t \beta_t}.$$

A proof of this claim is presented in the appendix A. Furthermore, when this step size condition is met, solutions to the discrete logistic equation of a given initial point are

bounded above by solutions with greater initial values and bounded below by solutions of lesser initial values. Existence and uniqueness guarantee this feature for the continuous model, but when comparing solutions to the discrete model, this is an important feature for solutions to have because it allows one to know for certain that one solution dominates another.

2.4 Logistic Approximation of Dynamics on Connected Graphs

Ultimately, for a mean-field representation, one would like to find a simple scalar equation that is a close approximation to (2.4), and determine what structural conditions must exist to allow such a scalar reduction. By approaching this question from the point of view of algebraic graph theory, we find that, for connected graphs, if A is either a doubly stochastic or a symmetric adjacency matrix, then the largest singular value can be used to find the closest rank-1 approximation to the original matrix in 2-norm, $\|\cdot\|_2$. Thus, by identifying the singular values of A, one can reduce the dimensionality of the system to a simple scalar approximation to (2.4). In general, graph topologies that permit doubly stochastic adjacency matrices are known to be contained in the family of strongly connected graphs, and we refer the reader to [41] for a more detailed technical discussion of this topic. The study of doubly stochastic systems is relevant for engineered systems, where the graph topology is constructed to have this doubly stochastic property.

The coordinated control of multi-agent systems [42] and the application of distributed consensus algorithms [43], for example, are often constructed with doubly stochastic communication topologies. Understanding the dynamics of robust information sharing amongst multi-agent systems is presently an ongoing area of research.

Though there exist matrices that are both doubly stochastic and symmetric, it is possible for a matrix to be doubly stochastic without being symmetric, or symmetric without being doubly stochastic. The latter case is more likely to occur naturally, but requires a relaxation of the row stochastic condition. Therefore, we will present the results for doubly stochastic matrices, followed by the results for symmetric matrices. We remark that the ability to utilize the matrix description of the network is critical for performing the scalar reduction in both cases, and we shall first review some important results from linear algebra that will be of use.

Suppose a given matrix A is symmetric, that is $A = A^T$. For $A \in R^{N\times N}$ and $A = A^T$, it is known that there exists an orthogonal matrix $W \in R^{N\times N}$ that diagonalizes A [44]:

$$A = WDW^T, \text{ with } D = diag\left\{\lambda_1, \ldots, \lambda_N\right\}, \tag{2.8}$$

where $\lambda_i \in R$ is the i^{th} eigenvalue of A such that $|\lambda_i| \geq |\lambda_{i+1}|$. Moreover, since the singular values of A are the positive square roots of the eigenvalues of A^TA, the result (2.8) and the following imply that each singular value of A is the absolute value of an eigenvalue of A:

$$A^TA = A^2 = WD^2W^T, \text{ such that } D^2 = diag\left\{\sigma_1^2, \ldots, \sigma_n^2\right\}$$

where σ_i is the i^{th} largest singular value of A [44].

Because of the close relationship between the eigenvalues and the singular values of real symmetric matrices, the problem of identifying the largest singular value is equivalent to the identification of the spectral radius, $\rho(A)$. One can appeal to the Perron-Frobenius theorem for row stochastic matrices to determine $\rho(A) = 1$ [37]. Denoting an N-dimensional vector of ones by 1_N, the row stochasticity of A implies that $A1_N = 1_N$ is an eigenvector of A with eigenvalue 1. The normalized eigenvector $1_N/\sqrt{N}$ is then the first column, w_1, of W in (2.8). Thus,

$$\sigma_1 = \lambda_1 = 1, \text{ and } w_1 = \frac{1}{\sqrt{N}} 1_N. \tag{2.9}$$

With the largest singular value and corresponding eigenvector identified, one is able to determine the closest rank-1 approximation in matrix 2-norm to an arbitrary adjacency matrix that is both doubly stochastic and symmetric. To see this, suppose $A = A^T \in R^{N\times N}$ and W is an orthogonal matrix that diagonalizes A as in (2.8). Then A can equivalently be represented as the series

$$A = \sum_{i=1}^{n} \lambda_i w_i w_i^T, \tag{2.10}$$

where λ_i is the i^{th} eigenvalue value of A and w_i is the i^{th} column of W. Furthermore, the closest rank-k approximation to A in matrix 2-norm is $X = \sum_{i=1}^{k} \lambda_i w_i w_i^T$, for $0 \leq k \leq \text{rank}\,(A)$. By recalling that the singular values of a symmetric matrix are the absolute value of its eigenvalues, a more general proof of this statement is provided in [45], with

the symmetry requirement relaxed and replacement of the Frobenius norm by the matrix 2-norm.

Using these properties of symmetric doubly stochastic matrices, one can rigorously define how well the scalar logistic model approximates the graph-based model, as proved in the appendix B. We now are able to define how well the scalar logistic model approximates the graph-based model for a doubly stochastic connected network topology from the following statement. Given a system of equations of the form (2.4) and defined on a doubly stochastic connected network with a reachable population of n members, the solution to the scalar logistic equation

$$\begin{cases} x_{t+h} = x_t + h\beta_t x_t \left(1 - x_t\right) \\ \\ x_0 = \frac{|p_0|_1}{N} \end{cases} \tag{2.11}$$

approximates the average value of the elements of p_t to an accuracy of order $h\sigma_2$.

When comparing this result to the direct solution of the completely connected case (2.5), it comes as no coincidence that the second largest singular value for the completely connected adjacency matrix is $(N-1)^{-1}$. To display this fact, equation (2.6) can be written in the form

$$x_{t+h} = x_t + h\beta_t x_t \left(1 - x_t\right) + \frac{h\beta_t}{N-1}\left(\frac{|p_t|_2^2}{N} - x_t^2\right).$$

We emphasize that the important feature of the network topology that produces this result is that the weighted adjacency matrix is doubly stochastic.

The error terms in these equations define a bound on the magnitude of fluctuations of mean-field solutions about the logistic solution at each step. The dependence of the

error term on the structure of the adjacency matrix alludes to the notion of structural convergence where the error converges to zero as the doubly stochastic or symmetric graph essentially becomes more completely connected in the sense of its matrix 2-norm. As a practical example, take an undirected cyclic graph of N nodes, where each node has k neighbors and each edge has a weight of $1/k$. The adjacency matrix of this system is both doubly stochastic and circulant, and its second largest eigenvalue is given by [46]:

$$\lambda_2 = \sum_{m=0}^{N-1} c_m e^{-i2\pi m/N}. \tag{2.12}$$

Because the exponential terms of (2.12) are symmetric about the real axis, the sum of imaginary terms is zero and the sum of real terms can be found by doubling the sum of the real terms over the interval $[0, \pi]$. Since $c_m = 0$ where edges do not exist and $c_m = 1/k$ where they do, one looks at the sum of the real terms to obtain the lower bound:

$$\cos\left(\pi \frac{k}{N}\right) \leq \lambda_2. \tag{2.13}$$

It is obvious that for fixed N, each edge added to the system by increasing k will make the system more completely connected. As this system grows, however, the lower bound λ_2 shows that the mean-field solutions will certainly not converge to the logistic solutions if the degree of each node does not increase at the same rate as the population.

Though one might perceive the double stochasticity requirement to be rather strict, requiring interaction symmetry between agents is quite realistic. For information spreading phenomena that involve direct one-on-one contact between members of a population, say during the spread of diseases or computer viruses, the amount of time two members

spend in an interaction is symmetric. When this amount of time is scaled by the total amount of time per period of interest, then one can conceivably obtain a symmetric non-negative weighted adjacency matrix whose row sums are between zero and one. In this case, the error term is defined by the largest singular value of the difference between the given adjacency matrix and the lowest rank approximation ($\text{rank}(A) = 1$) of a row stochastic matrix, denoted R_1, and whose elements are all N^{-1}. Similar to the procedure used to obtain (2.11) from (2.4), one begins with

$$p_{t+h} = p_t + h\beta_t \left(\mathrm{I} - diag\{p_t\}\right)\left(A + R_1 - R_1\right) p_t \tag{2.14}$$

to obtain

$$x_{t+h} = x_t + h\beta_t x_t \left(1 - x_t\right) + O\left(h \left\|A - R_1\right\|_2\right). \tag{2.15}$$

For example, let us consider chain of linked nodes arranged in a line such that each node has only two neighbors except for the nodes on the ends who each have just one neighbor. The adjacency matrix of this system will only have have nonzero elements on its upper diagonal, lower diagonal, or both. If symmetric interactions occur on this network, then one can apply (2.15) to this system. In this case, suppose each and every interaction takes place for the same proportion of a given time step, say $h/2$ so that each element along the upper and lower diagonal is $1/2$ and each row sum lies on the unit interval. As defined in [46], the structures of A and $(A - R_1)$ are both *banded Toeplitz matrices*, which are a class of matrices that asymptotically converge to their *circulant* analogs. Here, the

linear chain of linked nodes yields a circulant system, say A_C, by connecting the two ends of the chain. Hence, one can generally argue that for large chains of linked nodes, the results of (2.12) and (2.13) indicate that the mean-field solution for this system does not converge to a logistic solution in the thermodynamic limit. For even small linked chains of agents, such as $N = 10$, one finds that A_C is a sufficient approximation of A for logistic dynamics since $\|A - R_1\|_2 - \|A_C - R_1\|_2 \leq 1.5 \times 10^{-3}$, and $\|A - R_1\|_2 = 0.9995$ indicates that the mean-field behavior of the linear chain is not expected to be logistic.

2.5 Mean-Field Behavior of Heterogeneous Networks

Thus far we have discussed the accuracy of logistic mean-field solutions as approximations to actual solutions of information propagation on the network. Conversely, an important question to address is how well the logistic solution approximates mean-field behavior for arbitrary heterogeneous networks that perhaps do not have doubly stochastic or symmetric edge weights, or whose structure is not defined algebraically. How can one analyze the influence of a network's structural properties on the dynamics that occur on the network? One approach to answering this question involves comparing the parameters that describe the graph structure to the parameters that describe the dynamics on the the graph. For the parameters used to describe the SI type dynamics, one can choose the transmission rate and the initial value of a scalar logistic approximation to the ensemble average of mean-field solutions.

Since a discrete logistic solution of (2.1) is determined by the transmission rate β_t and an initial condition that depends on population size, one should be able to deduce a β_t and initial value for a given time series that resembles a discrete logistic solution. Once these are known, one can infer a corresponding homogeneous network whose mean population behavior produces an almost identical logistic solution. Therefore, if the mean behavior of a heterogeneous network is known, then one can describe similar system dynamics in terms of a homogeneous network by fitting a discrete logistic solution to the mean heterogeneous solution. Here, the error of the approximation is defined as the 2-norm of the difference between the heterogeneous solution data points and points of the discrete logistic approximation for the first 100 time steps.

The mean behavior of a realization where only one agent is informed depends on the size of the reachable set for that initially informed agent. In general, the reachable set is the union of the reachable sets of all initially informed individuals. Thus, when comparing the mean behavior for different initially informed node sets, one should be sure that their reachable sets are of the same size. It is often useful to identify a set of strongly connected nodes since each node contained in a strongly connected set must necessarily have the same reachable set of nodes [37]. To keep comparisons simple, one can compute the mean population behavior with respect to time when only one node is initially informed, and repeat this computation for each node in the strongly connected set. We explore this idea for a graph topology defined by a naturally occurring scale-free

graph, a family of Watts-Strogatz graphs, and a family of linked sub-graphs where each sub-graph is itself a Watts-Strogatz graph.

2.5.1 Scale-free graph example

Here, we begin with a graph topology defined by a network of Wikipedia administrator voters [47], an example of a naturally occurring directed social network with uncorrelated degree distributions. The in-degree distribution is approximately power law distributed $prob(k) = 0.293 * k^{-1.357}$, and the sample correlation coefficient between in-degree and out-degree is $\gamma = 0.387$. Here, the sample correlation coefficient γ between two finite data sets, say x and y, is calculated according to [48]:

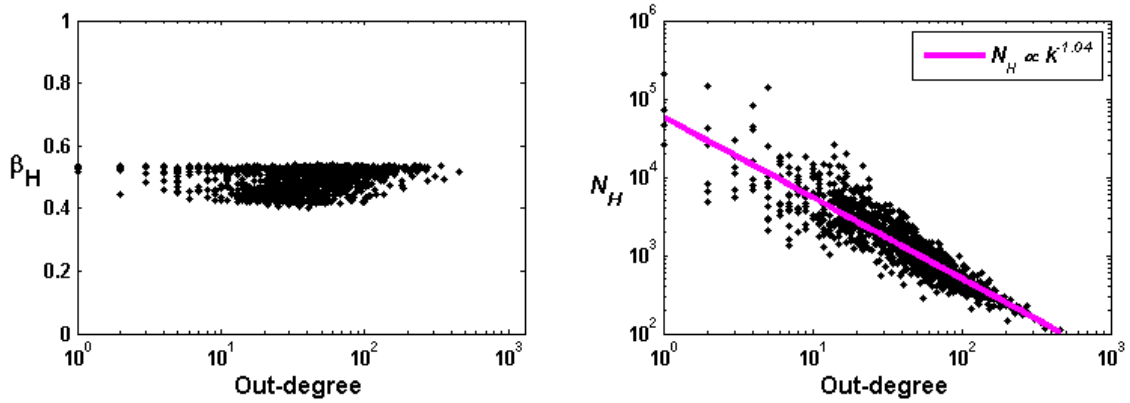
$$\gamma = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{2.16}$$

where $\bar{x}$ and $\bar{y}$ represent the mean values of the x and y data sets, respectively.

In the context of information propagation, suppose agents i and j are neighbors. If i votes for j, then this indicates a directed relationship where we know i at least pays attention to j. In this sense, information is understood to flow from j to i and indicates the presence of a directed edge (j, i). For this study we compared the realizations for each node in the largest set $\{V_{SC}\}$ defined as the set of strongly connected nodes that contains the node of greatest out-degree. The set $\{V_{SC}\}$ contains 1300 nodes and a reachable set of 5158 nodes.

Denoting k_i as the number of edges directed towards agent i, each node is assumed to follow his in-neighbors equally such that each edge directed towards agent i has the

value $1/k_i$, which shall be referred to as the *unbiased weighting scheme*. Using the unbiased weighting scheme allows our study to focus on network structure by controlling for edge weight. The dynamics were simulated according to (2.4) for the Wikipedia voting adjacency matrix, and the results are shown in Figures 2.1 and 2.2. The realizations were generated using $\beta = 1$, to control for transmission rate, and the average probability of being informed was calculated over the entire reachable set at each time step for 100 steps with $h = 0.99$. For each node in $\{V_{SC}\}$, a realization was computed where the given node has an initial probability of 1 and all other nodes have initial probabilities of 0.



(a) (b)

Figure 2.1: For each node contained in $\{V_{SC}\}$, data was simulated according to (2.4), where each given node is the only one initially informed. A discrete logistic solution was then fit to each initial node's mean-field solution. (a) and (b) depict the optimal parameter values that minimize the 2-norm difference between the original mean-field solution and the approximate logistic solution for each initial node, and plotted with respect to the initial node's out-degree (k). The logistic approximations have a mean 2-norm difference of 0.0326 with a 0.0024 standard deviation, and range of $[0.0277, 0.0413]$.

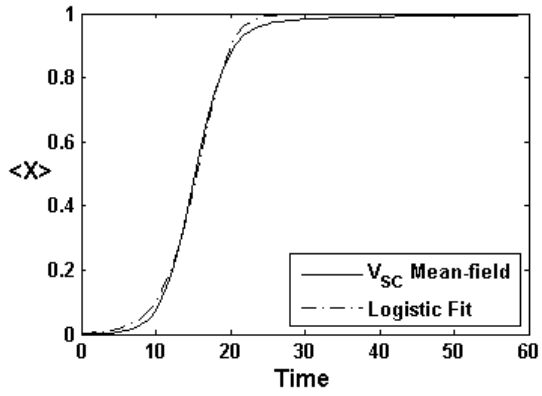


Figure 2.2: Logistic approximation of the ensemble over all realizations for the set $\{V_{SC}\}$. The approximation has $\beta_H = 0.4801$, $N_H = 340$, and a 2-norm error of 0.0624.

Since the transmission rate used to generate each realization was a constant value, we took it as an assumption that the best fit scalar logistic approximation is generated by an unknown constant transmission rate, β. Given the mean behavior of each heterogeneous realization, we first identified the best fit transmission rate (β_H) using a least squares method because β is linear with respect to the dynamics at each time step. We then identified the optimal initial condition using an iterative process. By assuming also that only one individual is initially informed, one can deduce an effective homogeneous population size (N_H) from the initial value of the logistic fit since the initial value is the inverse of the homogeneous population size in this case.

The closest approximate logistic dynamics are depicted in Figure 2.1 in terms of the β_H and N_H parameter values for the set $\{V_{SC}\}$. The mean-field solutions have β_H values that appear to be rather consistent regardless of the initial node's out-degree, as shown in Figure 2.1a, while Figure 2.1b suggests that the values of N_H depend logarithmically on initial node out-degree. Figure 2.2 shows how well a discrete logistic solution describes the

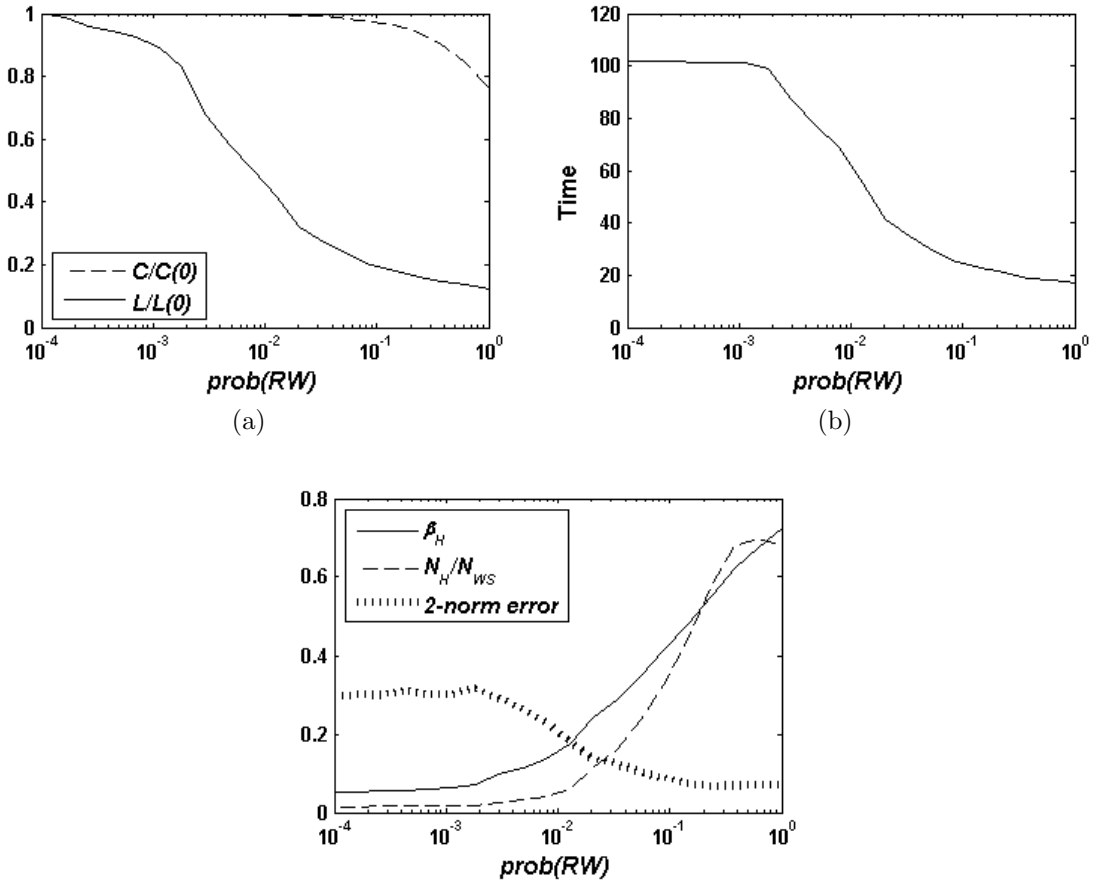
ensemble average of population mean-field solutions for the set $\{V_{SC}\}$. The best fit logistic approximations have 2-norm errors that lie outside the range of 2-norm errors for their individual mean-field solutions, (i.e. $0.0624 \notin [0.0277, 0.0413]$). Hence, each individual mean-field solution is closer to having discrete logistic behavior than the expected average behavior of the whole system.

By approximating mean-field solutions on an arbitrary network by that of a homogeneous network, one can interpret the homogeneous network size N_H as being a descriptor of the ease with which information can spread through the network. This argument is particularly convincing in cases when the values of β_H are essentially the same over all mean-field solutions, because then the bounding of solutions depends only on the initial values, and thus N_H, when the step-size condition is met, as discussed in the appendix A. When only one node is initially informed, large homogeneous networks will naturally take longer for information to diffuse through than relatively smaller homogeneous networks. Therefore, one can compare a given node's effect on the the network to that of other nodes of the network. For the unbiased weighting scheme, these results do not contradict the findings of [8,18] since faster rates of information diffusion in our model are correlated with higher out-degree of the initially informed node. The correlation value between an initial node's $\log_{10}(k_{out})$ and its $\log_{10}(N_H)$ is $\gamma = -0.9001$.

2.5.2 Watts-Strogatz Graphs

One can control for structural effects on the dynamics caused by the average degree of nodes on a network by analyzing a family of networks originally studied by Watts and Strogatz [1], where a regular undirected graph is constructed such that each node has the same degree and a subset of the edges are "rewired" according to a given rewiring probability. We generated, for each rewiring probability ($prob(RW)$), a set of 20 Watts-Strogatz (WS) graphs of 1000 nodes and average degree of 20. Each graph was given an unbiased edge weighting scheme, and the mean-field solutions were generated with a transmission rate of $\beta = 1$. The logistic solutions were approximated following the procedure described in Section 2.5.1 for which each node is initially informed with probability 1 and all else zero. For a set of rewiring probabilities ranging from $prob(RW) = 0$ to $prob(RW) = 1$, the average graph structural parameters are shown in Figure 2.3a, while the average homogeneous approximation parameters are shown in Figures 2.3b and 2.3c. It is noted that the trends of L and C in Figure 2.3a suggest the presence of small-world structures logarithmically centered about $prob(RW) = 0.1$, where the ratio of C to L is greatest [1].

Figure 2.3c shows that the accuracy of the homogeneous approximation improves as the graph becomes more random. It is also noted that both β_H and N_H increase as $prob(RW)$ increases. The increase in N_H caused by an increase in $prob(RW)$ seems counterintuitive because one would ordinarily expect a network of shorter average path length to seem smaller from the perspective of the information diffusing on the network.



(a) (b) (c)

Figure 2.3: Similar to the original study conducted by Watts and Strogatz [1], 20 WS graphs of size $N_{WS} = 1000$ were generated and their graph parameters were averaged at each rewiring probability. (a) WS graph structure in terms of the average clustering coefficient (C) and average characteristic path lengths (L) over all nodes, as defined in Section C of the appendix. Both C and L are normalized with respect to their values for zero rewiring probability. (b) Average time steps for mean-field solutions of (2.4) to reach $x = 0.99$. (c) Average transmission rate (β_H), homogeneous population size (N_H), and 2-norm homogeneous approximation error. N_H is normalized with respect to the population size of the original WS network.

Parameter	Correlation
β_H	-0.9178
N_H	-0.8478
C	0.6304
L	0.9876

Table 2.1: Correlation coefficients calculated according to equation (2.16) over the spectrum of rewiring probabilities, relating the average parameters in the left column to the average number of steps needed for mean-field solutions to reach $x = 0.99$.

However, the value of β_H also increases along with $prob(RW)$, which likely counteracts this effect. It is also noted that the result of equation (2.13) for regular graphs indicates an order of accuracy that is proportional to $\cos(\pi(20)/(1000))$ for this case. The average homogeneous approximation errors of the scale-free graph of Section 2.5.1 is an order of magnitude smaller than those of the family of WS graphs, even though the population size of the scale-free graph is almost an order of magnitude larger. It is observed that the average error decreases with the value of $k/N = 0.02$ held constant during these simulations, which indicates that the mean-field behavior of random graphs is in this sense relatively more logistic than that of regular graphs. Here, we shall adopt the $\langle\cdot\rangle$ notation to denote the average value of a given parameter over all nodes at a fixed rewiring probability.

To determine which types of networks spread information the fastest, one can compare the average time it takes the system to reach 99% information saturation (*i.e.*, $\langle x\rangle = 0.99$) since in some cases it is possible to only reach 100% saturation in infinite time. The number of time steps needed for the system to reach 99% information saturation is

depicted in Figure 2.3b. As Table 2.1 suggests, the characteristic path length is a strong indicator of the rate at which information is able to diffuse through a WS network, while β_H has more of an effect on time needed to reach saturation than N_H. It is noted that the least amount of time needed to reach 99% saturation occurs for the set of graphs having rewiring probability $prob(RW) = 1$ (random graphs), and occurs in 8 fewer time steps on average than graphs of $prob(RW) = 0.1$ (small world graphs). This suggests that random networks spread information faster than small world graphs when controlling for average node degree.

2.5.3 Chain of Watts-Strogatz Graphs

To extend the analysis of WS graphs, suppose a network is constructed as a sequence of WS networks such that only one undirected edge connects two neighboring WS sub-networks, as depicted in Figure 2.4a. Applying the homogeneous approximation to this type of system allows comparison to both WS graphs and chain graphs on the macro scale, while also being able to probe the behavior of individual nodes, such as those that connect the distinct WS subgraphs, on the local scale.

For the chain of WS graphs, the average graph structural parameters are shown in Figure 2.5a, while the average homogeneous approximation parameters are shown in Figure 2.5b. Although the trend of C in Figure 2.5a for the chain of WS graphs is almost identical to that of Figure 2.3a for a single WS graph, there is a noticeable difference in the trends of L among Figures 2.3a and 2.5a with respect to $prob(RW)$. One might

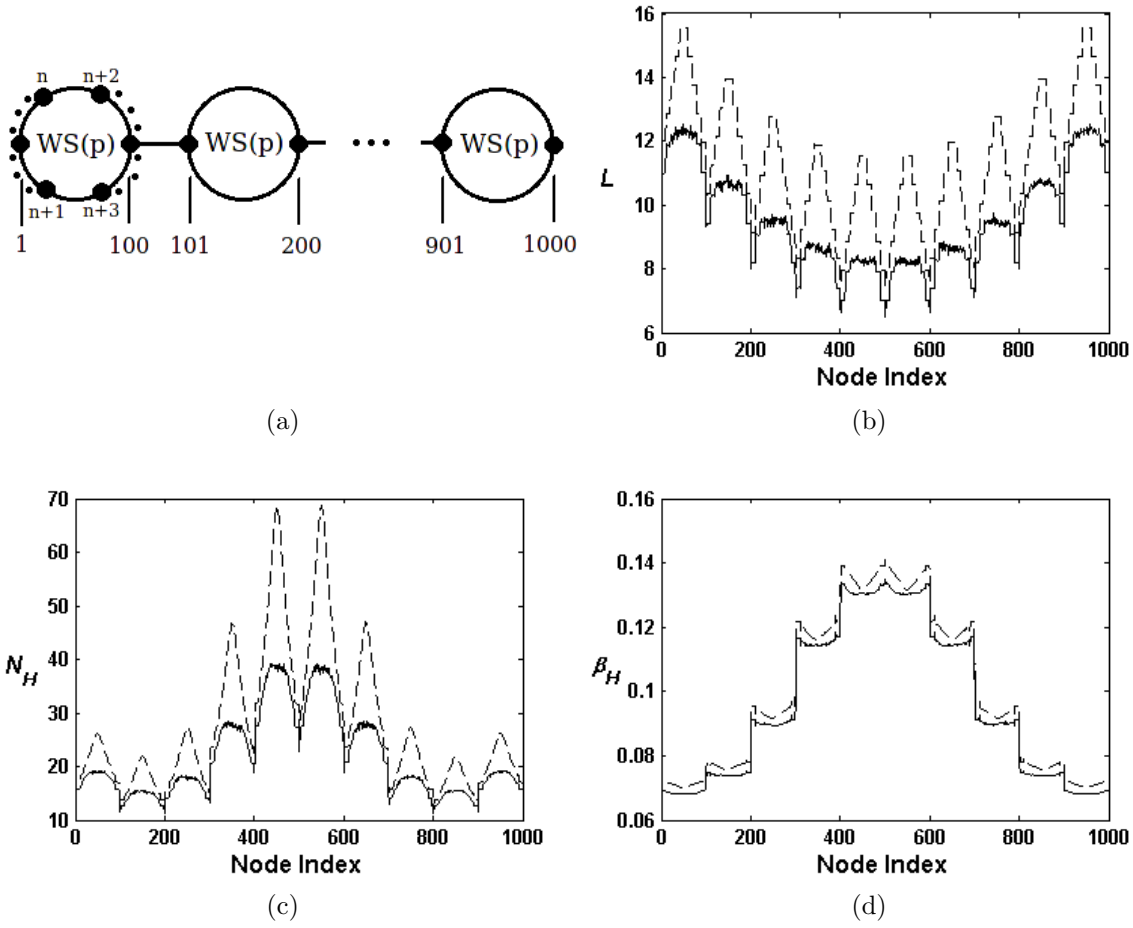
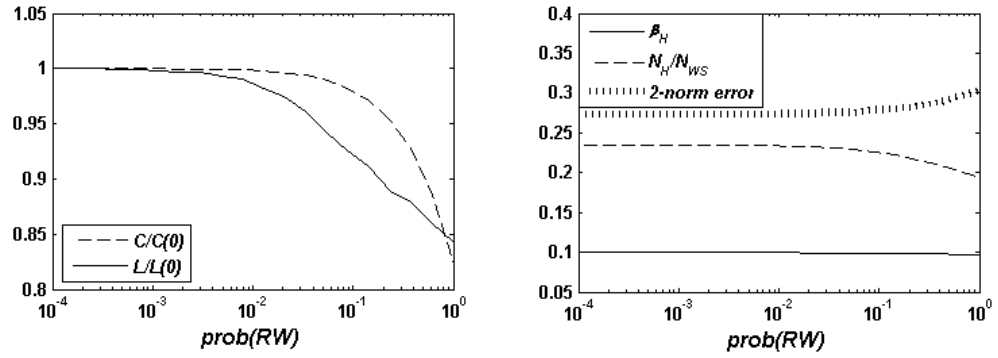


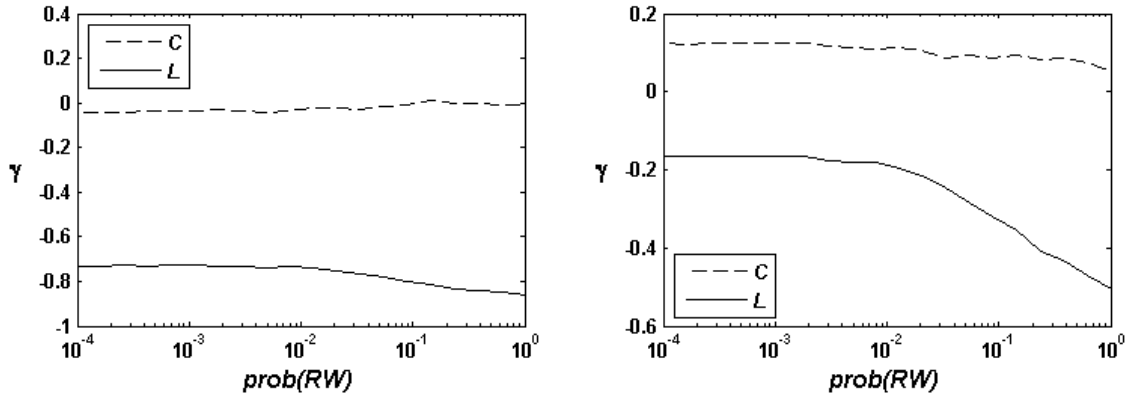
Figure 2.4: (a) Chain of 10 WS graphs with 100 nodes each are linked together with one edge connecting each WS graph. The nodes are labeled left to right, and alternating top to bottom, with increasing index. (b) Average C with respect to individual node index. (c) Average N_H with respect to individual node index. (d) Average β_H with respect to individual node index. In (b), (c), and (d) the dashed line represents data for $prob(RW) = 0$, and the solid line represents data for $prob(RW) = 1$.



(a) (b)

Figure 2.5: Similar to the original study conducted by Watts and Strogatz [1], 20 chains of ten WS graphs of size $N_{WS} = 100$ were generated and their graph parameters were averaged at each rewiring probability ($prob(RW)$). (a) WS graph structure in terms of the average clustering coefficient (C) and average characteristic path lengths (L) over all nodes. Both C and L are normalized with respect to their values for $prob(RW) = 0$. (b) Average transmission rate (β_H), homogeneous population size (N_H), and average 2-norm error. N_H is normalized with respect to the population size of the constituent WS networks.

hypothesize that the discrepancy of L between the two systems can be attributed to the fact that each WS sub-graph of the WS graph chain has 100 members instead of 1000. However, the identical behavior of C for the two systems suggests that the chain structure of the WS sub-graphs has a more significant impact on L with respect to $prob(RW)$ since rewirings were not allowed to occur between each WS sub-graph. When the average homogeneous approximation parameters are compared, one finds that the data of Figure 2.5b show trends that oppose those of Figure 2.3c: Figure 2.5b shows an increasing error and barely decreasing β_H and N_H as $prob(RW)$ increases. The reason for the opposing trend in the average homogeneous data for the chain of WS graphs versus the single large WS graph is that the WS sub-graphs are linked as a sequential chain.



(a) (b)

Figure 2.6: (a) Average correlation values between C and L, and β_H for the chain of WS graphs. (b) Average correlation values between C and L, and N_H for the chain of WS graphs.

As rewiring probability approaches $prob(RW) = 1$, each WS sub-graph becomes better mixed and appears to behave more as one entity since information diffuses fastest on the single WS network level, as shown in Figure 2.3b. While the WS sub-graphs become more mixed, the increase in homogeneous approximation error is likely explained by the fact that chains of nodes do not produce logistic mean-field solutions, as discussed in Section 2.4.

Figure 2.6 shows that the average clustering coefficient is weakly correlated with both β_H and N_H, while the characteristic path length is somewhat correlated with β_H over all rewiring probabilities and becomes more correlated with N_H as rewiring probability increases. As the characteristic path length decreases, the negative correlation with both β_H and N_H indicates that the information not only diffuses faster, but through an effectively larger network. Hence, the structural effects that cause an increase in β_H

values offset those that cause a decrease in N_H, and results in an average of 76 time steps to reach 99% of saturation for each rewiring probability. In this case, one can interpret the networks as being equally capable of diffusing information.

The expected mean-field solutions are actually quite similar in performance to each other for this type of system since the collection of ensemble averages of the mean-field solutions over the spectrum of rewiring probabilities have an average 2-norm error of 0.0624. Compared to the error of the homogeneous approximation to mean-field solutions, which has an average value of 0.2778 in 2-norm, as deduced from the data of Figure 2.5b, the homogeneous approximation is still sensitive enough to detect subtle features in the mean-field solutions despite how non-logistic the mean-field solutions are.

At the individual node level, Figure 2.4b shows the characteristic path lengths at each end of the rewiring probability spectrum, along with their corresponding index labels. At this level of detail, it is easy to see the effects that the sub-graph connecting nodes have on the dynamics relative to their global location on the graph. By comparing Figure 2.4b to 2.4c and 2.4d one is able to observe how the characteristic path length of each node is reflected by its effective homogeneous network size and effective transmission rate, respectively. Figure 2.4c shows how the sub-graph connecting nodes perceive the smallest effective homogeneous networks, while those towards the center of the network perceive the largest effective homogeneous networks of all. When this observation is compared to the average characteristic path length of each node, as observed in Figure 2.4b, one

finds this observation to, again, be a counterintuitive result that can be explained by the opposing effect of β_H as seen in Figure 2.4d.

2.6 Conclusions

By focusing our attention on SI dynamics, we have shown the importance of applying algebraic graph theory to dynamic processes in a simple information spreading context, and the new physical insight it is able to provide to information spreading phenomena. In contrast to the application of approximate parameter distributions to the dynamic equations, such as power-law degree distributions, adjacency matrices preserve the exact global structure of a weighted network. Here, we were able to also use adjacency matrices to rigorously attribute the size of fluctuations about the population-level solutions to the structural similarity between a given graph and a completely connected graph. In the case of completely connected graphs, the fluctuations were found to be attributed to finite size effects.

Specifically, we have constructively shown that the agent-based and scalar logistic models are in exact agreement for the completely connected case in the limit as the number of agents in the system approaches infinity, as conjectured previously in the different research communities. For homogeneous systems consisting of a finite number of agents, the singular values of the graph adjacency matrix produce the closest logistic approximation to the completely connected agent-based dynamics. This result was extended to connected networks that are doubly stochastic or with symmetric interactions so that

systems of this type can generally be approximated by the discrete logistic equation. Although double stochasticity is typically only relevant to engineered systems, we have seen that our analytic methods are applicable to naturally occurring systems since propagation mechanisms with interaction symmetry are quite common. We also discussed how one can analyze the logistic behavior of arbitrary heterogeneous network topologies.

Moreover, by analyzing average population behavior, we found that there are instances when solutions to heterogeneous dynamics of one set of parameter values appear to be well approximated homogeneous dynamics for a different set of parameter values. If the entire network structure and set of parameter values are known, an implication of being able to use homogeneous systems to approximate heterogeneous systems is that it provides a standard way of comparing the dynamics of two heterogeneous systems. In general, the logistic behavior of any two homogeneous networks can be compared to each other. To avoid results that may be misleading, however, we advocate only comparing networks of the same size when making network versus network comparisons and comparing nodes with the same size reachable sets when making node versus node comparisons.

In regards to the inverse problem of using observable system behavior to infer graph features, coarse descriptions, such as mean-field behavior, are likely to not contain enough detail to distinguish one graph topology from another. Caution should be exercised when observing similar mean-field behavior of different logistic dynamical systems because uniqueness properties relating mean-field behavior to heterogeneous graph structure have

yet to be established, particularly if they have different weighting schemes. Just as there can be two completely different graph structures that produce the exact same mean-field behavior, there could also be two identical graph structures of different weighting schemes that could exhibit different mean-field behaviors.

Chapter 3

Population behavior on topic backbones

Suppose now that in addition to having a static network structure, we also have time series data on parcels of information that are propagating on this network. As discussed in Chapter 2, when analyzing the effect of network structure on the spread of information, we first want to find the irreducible subnetwork on which our parcel of information exists. We can then extract the population-level behavior of this parcel, and quantify the fluctuations about the mean-field population behavior.

In this chapter we introduce a large data set of user contributed posts to the Twitter social network, which contains approximately 467 million posts among 42 million users. We also define *topic backbone* structures of the network that are able to capture the network's cascade edges (edges where an adoption event is observed), yet each have multiple orders of magnitude fewer nodes than the full network. The topic backbones are tenable in size and implicitly capture the core connected component of the network, thus

	SNAP (users=42M,tweets=467M)		
Topic	Hashtags	Users	Uses/HT
Business	27	20k	1,155
Celebrities	32	26k	1,009
Politics	485	349k	2,020
Sci/Tech	33	415k	6,889
Sports	98	76k	3,274

Table 3.1: Statistics of the SNAP data set.

allowing us to begin applying the tools developed in Chapter 2 to real social network data.

3.1 Datasets

We chose Twitter to analyze user behavior via our genotype model since Twitter has millions of active users and messages have a known source, audience, time stamp and content. Similar analysis can be performed in other social media networks with a known follower structure and knowledge of the shared content (memes, URLs or buzz-words) in time.

3.1.1 Twitter follower structure and messages

We use a large dataset from Twitter, SNAP [49], which includes a 20% sample of all tweets from June to December 2009 and contains 467 million posts. The complete follower structure [50] for the Twitter social network structure during this time frame is based on the complete follower crawl of Kwak et al. [50], and includes over 42 million

Twitter users and over 1.47 billion edges. The SNAP data set's statistics are summarized in Table 3.1 with sample sets of hashtag topics.

3.1.2 Grouping hashtags into topics

A hashtag is a user generated token, usually written as a string of characters following a pound sign ("#"), that annotates a message and allows users to participate in global discussions [51]. While hashtags present a concise vocabulary to annotate content, they are free-text user-defined entities. Hence, we need to group them into topics in order to summarize network behavior at the topic level, which will help reduce the massive Twitter network to a few relevant reachable sets. The intention here is to eventually apply the agent-based network model to the Twitter network, which can only be done if the Twitter network is reduced to a scale that is both computationally tractable, and contains all of the relevant transmission edges for each hashtag topic set.

In this work we assume each hashtag belongs to exactly one topic*, while in a more general framework disseminated hashtags (or URLs, memes, etc.) can be "softly" assigned to more than one topic. We work with five general topics: Sports, Politics, Celebrities, Business and Science/Technology. We obtain a set of 100 hashtag annotations from a recent work by Romero and colleagues [52], further augmented by a set of curated business-related hashtags [53]. We combine this initial set of annotated hashtags with a larger set based on text classification.The set of manually curated hashtags from

*With limited exception, as noted in Appendix D

previous work is modest compared to the size of content disseminated in a large system like Twitter. Hence, sparsity of each hashtag usage is a limiting factor in characterizing topic-specific behavior.

To increase the number of considered hashtags, we adopt a systematic approach for annotating hashtags based on URLs within the tweets. To associate tweets with topics, we treat user-generated hashtags as tokens that carry topical identity, similar to previous studies [52]. Users include hashtags to annotate (topically) their tweets and to participate in a specific community discussion [49]. Adopting the appropriate hashtag for a message ensures better chances of surfacing the content in search as well as attracting the attention of interested followers.

We pair non-annotated hashtags with web URLs, based on co-occurrence within posts. We extract relevant text content from each URL destination (most commonly news articles from foxnews.com, cnn.com, bbc.co.uk) and build a corpus of texts related to each hashtag. We then classify the URL texts in one of our 5 topics using the MALLET [54] text classification framework trained for our topics of interest. In order to train the MALLET topic classifier we use annotated text from two widely used topic-annotated text collections: the 20 newsgroups dataset [55] and the News Space [56]. Additional ground-truth text collections can be used for wider topic coverage and to improve the accuracy.

As a result, we get a frequency distribution of topic classification for frequent (associated with at least 5 tweets) hashtags. The topic annotation of the hashtag is the

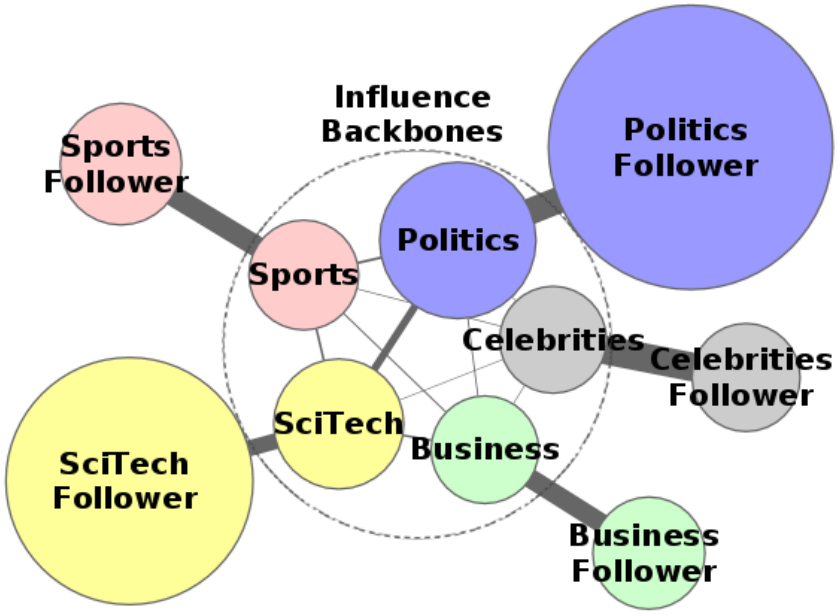


Figure 3.1: Overlap among topic influence and corresponding follower subnetworks (in SNAP). Each network is represented as a node, with every topic represented by an influence (encircled in the middle) and a follower network. Node sizes are proportional to the size of the network (ranging from $120k$ for Celebrities to $42m$ for Politics Follower). Edge width is proportional to the Jaccard similarity of the networks (ranging from 10^{-3} inter-topic edges to 10^{-1} between corresponding influence-follower networks).

topic of highest frequency. The number of hashtags and their usage statistics in our final topic-annotated set are presented in Table 3.1 (columns Users and Uses/HT), and the specific hashtags used in this study are contained in Appendix D.

3.2 Topic-specific influence backbones

Directed Twitter links do not necessarily represent friendship ties but sometimes merely interest in the information produced by the followee. This leads to a denser link structure than in traditional social networks. As such, a follower network provides a middle ground between traditional broadcast media distribution (some nodes represent

media outlets with millions of followers) and a more personal information exchange. Recent research has demonstrated that many follower links are actually reciprocal [57], suggesting that a significant portion of the network actually corresponds to personal friendship ties. On the other hand, there are a number of extremely high fan-out nodes corresponding to media outlets, companies and prominent public figures. As a result, it is difficult to judge how individual influence propagates in the network by simply observing the network structure on its own. Instead this task requires understanding of the behavior of nodes.

With regards to population-level dynamic behavior on a network, the spread of information on a network has been primarily explored using models adopted from epidemiology [17, 33], and have been applied to describe propagation rates of memes (i.e., Twitter hashtags) in social media [58]. We adopt these methods of analysis to evaluate the population-level topic behavior on influence networks, and assume a simple contagion model as the underlying propagation process in our data sets.

By observing the behavior of agents (adoption, reposting, etc.) one can reveal the underlying backbones along which topic-specific information is disseminated. In this section, we study the propagation of hashtags within Twitter to identify *topical influence backbones* — sub-networks that correspond to the dynamic user behavior. We superimpose the latter over the static follower structure and perform a thorough comparative analysis to understand their differences in terms of structure and population-level user behavior.

3.2.1 Influence backbone definition and structure

An *influence edge* $e_i(u, v)$ connects a followee u who has adopted at least one hashtag h within a topic T_i before the corresponding follower v. Hence, the influence network $N_i(U, E_i)$ for topic T_i is a subnetwork of the follower network $N(U, E)$ (including the same set of nodes U and a subset of the follower edges $E_i \in E$). To measure the importance of each edge, we weight the edges of the influence network by the number of hashtags adopted by the followee after the corresponding follower, and within the same topic. According to the notation defined in Chapter2, the Twitter network $N(U, E)$ represents a graph $G(V, E)$. Each user on the Twitter network represents a vertex (node) on the graph such that $U = V$.

First, we seek to understand the differences between the influence backbones and the static follower network. Figure 3.1 presents the overlap among influence backbones and their corresponding follower network. For this comparison, we augment an influence network with all follower edges among the same nodes to obtain the corresponding follower network. In the figure, each network is represented by a node whose size is proportional to the network size (in edges). Connection width is proportional to the *Jaccard Similarity (JS)* (measured as the relative overlap $|E_i \bigcap E_j|/|E_i \bigcup E_j|$) of the edge sets of the networks. The Jaccard similarity for influence and follower networks varies between 0.16 for Sports to 0.3 for Celebrities. The influence networks across topics do not have high overlap (JS values not exceeding 0.01), with the exception of Sci/Tech and Politics with $JS = 0.07$. This may be explained partially by the fact that these are the largest influ-

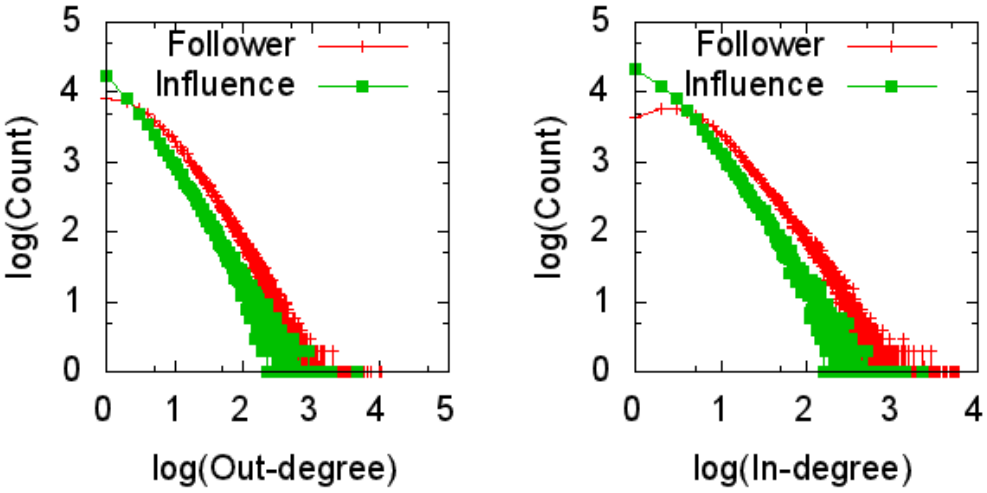


Figure 3.2: Out- and In-Degree distributions for the Follower and Influence networks for *Sports* for the SNAP dataset.

ence networks (5 and 11 million edges respectively). Another reason could be that there are some "expert" nodes who are influential and active in both topics.

The degree distributions of influence and follower networks within a topic maintain a similar shape. Figure 3.2 shows the in- and out-degree distributions for the *Sports* networks in the SNAP dataset. The most dramatic change in the distributions is for small degrees with almost one magnitude increase of the nodes of in-degree 1. Users who retain only a few influencers tend to have a variable number of followees, hence the in-degree distribution decreases for the whole range of degrees.

Beyond network sizes and overlap, we also quantify the structural differences of the influence backbone in terms of connected components. A *strongly connected component (SCC)* is a set of nodes with directed paths among every pair, while in a *weakly connected component (WCC)* connectivity via edges regardless of their direction is sufficient. Fig-

ure 3.3 compares the sizes of the largest SCC and WCC in the topic-specific networks as a fraction of the whole network size. When ignoring the direction (i.e. considering WCC), both the influence and follower structures have a single large component amounting to about 99% of the network. The communities that are active within a topic are connected, showing a network effect in the spread of hashtags, as opposed to multiple disjoint groups which would suggest a more network-agnostic adoption. When, however, one takes direction into consideration (SCC bars in Figure 3.3), the size of the SCC reduces drastically in the influence backbones. Less directed cycles remain in the influence backbone, resulting in a structure that is close to a directed acyclic graph with designated root sources (first adopters), middlemen (transmitters) and leaf consumers. The reduction in the size of the SCC is most drastic in the Celebrities topic, indicative of a more explicit traditional media structure: sources (celebrity outlets or profiles) with a large audience of followers and lacking feedback or cyclic influence.

We next address the issue of how a user's importance changes. In Figure 3.3 (bottom) we show the correlation of node ranking based on number of followers, followees and PageRank [59] in the influence and follower networks. The correlation of each pair of rankings is computed according to the *Kendall* τ rank correlation measure. The correlation is below 0.5 for all measures and topics. Global network importance (PageRank) is the most distorted when retaining only influence edges (0.4 versus 0.5 on average), while locally nodes with many followers (or followees) tend to retain proportional degrees in the influence network.

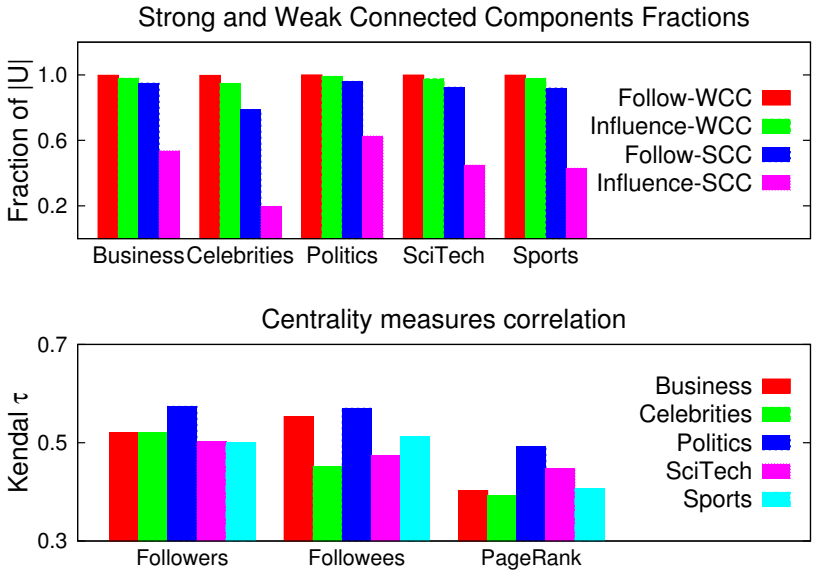


Figure 3.3: Largest weakly and strongly connected component (WCC and SCC) sizes as a fraction of the network size (top); and Kendall τ rank correlation of node importance measures for the influence and follower networks (bottom) for the SNAP dataset.

While the follower structure features a lot of reciprocal (bi-directional) links (above 50% on average), these reciprocal links disappear almost completely in the influence backbone (retaining 4% on average), as shown in Figure 3.4. This effect is most prominent in the Celebrities topic where reciprocal links drop from 36% to less than 1% in the influence network. Reciprocal links are related to friendship ties, i.e. nodes who are possibly friends declare interest in each other's posting by a bi-directional link. When it comes to influence, however, the ties tend to be uni-directional with only one of the nodes affecting the other.

Our comparative analysis of the influence and follower structure demonstrates that the influence backbone is quantitatively different from the overall follower network. The explanation for this lies in the fact that the influence backbone is based on the dynamic

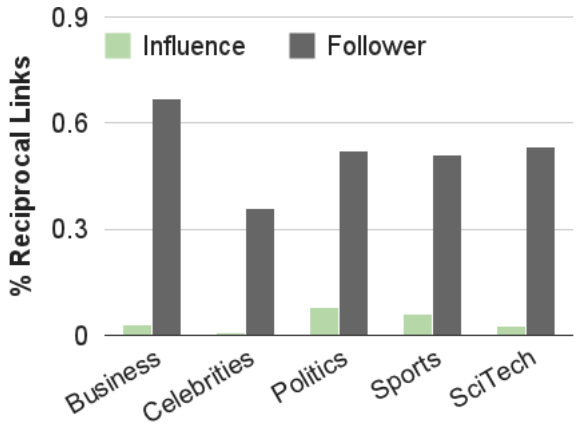
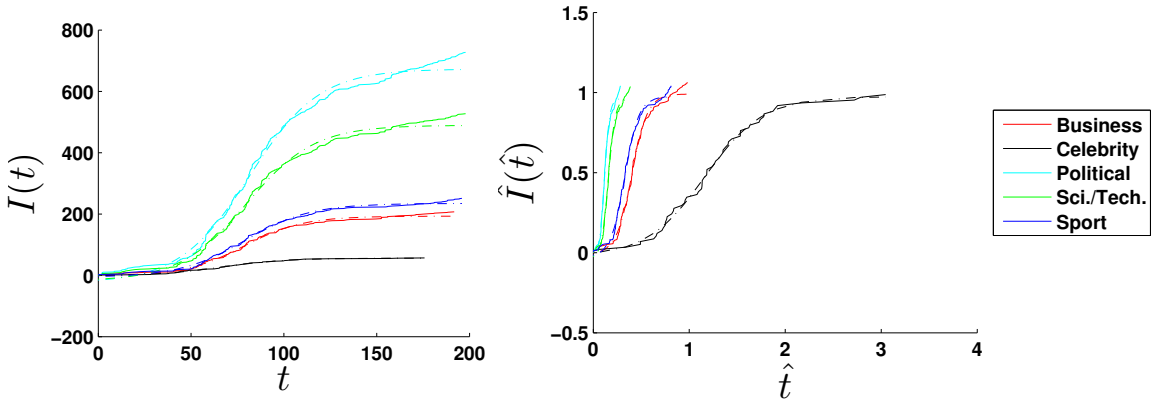


Figure 3.4: Comparison of the percentage of reciprocal (bi-directional) links in the influence and follower networks.

behavior of users (information dissemination on specific topics), while the follower structure represents the static topic-agnostic media channels among users. Not all followees tend to exert the same amount of influence over their audiences in the actual information dissemination process, giving rise to distinct topic-specific influence backbones.

3.2.2 Population behavior on topic backbones in Twitter

Thus far, the topical influence backbone networks are comprised of the individuals within a given topic. Since many users are members of more than one backbone, yet may be more responsive towards one topic than another, an ensuing question is whether dynamics on the topic backbones are consistent with individual behavior. Does the Business backbone, for example, propagate business hashtags faster than, say, the Sports backbone? In general, we find this hypothesis to be true, assuming that the underlying hashtag propagation process follows a simple epidemic-inspired compartmental population model.



(a) (b)

Figure 3.5: Example of typical regression result, from data of the Political hashtag *#beck*, referring to the political commentator Glenn Beck. (a) The measured data (solid lines) and the approximated regression function (dashed lines) in the unnormalized coordinates, and (b) the same data in the normalized coordinates. The plotted curves are colored according to the topic backbone that the *#beck* hashtag was detected on.

Compartmental population models are often implemented to study average behavior of a disease or meme within a population [17, 33, 58]. As discussed in Chapter 2, the simplest case is where we have only two classes of individuals, susceptible (S) and informed (I), a susceptible individual can become informed of a meme, and once informed will remain informed. Such coarse two-state models for simple contagions (i.e., cascades) describe average rates of adoption from one class of individuals to the next. For static populations, where $S + I = N$ for some fixed population of size N, the dynamics of a typical S-I process are defined by Hethcote [17] as:

$$\frac{dI}{dt} = \beta I(1 - \frac{I}{N}), \tag{3.1}$$

which has the solution

$$I(t) = \frac{NI(0)e^{\beta t}}{N + I(0)\left(e^{\beta t} - 1\right)}, \tag{3.2}$$

where β is the transmission rate and $I(t)$ is the size of the infected population at time t.

One can quantify and compare the contagiousness of a hashtag on different networks by comparing its respective β values. An example set of realizations is depicted in Figures 3.5a and 3.5b. It is important to note the sigmoidal shape of the adoption curves and their least-squares approximations. This sigmoidal shape is characteristic of the processes governed by (3.2).

For this particular study, we track a hashtag of known topic on the Twitter network in order to observe whether or not the hashtag is most *viral* on its own topic backbone. We begin by considering only hashtags that have been tweeted by users who are members of more than one topic backbone within the SNAP dataset. A distinct realization of (3.1) for a hashtag is defined by the total population of individuals who have tweeted that hashtag with respect to time.

When comparing the model defined by (3.2) to temporal hashtag data, one needs to account for the fact that the hashtag may have existed on the network prior to the time of initiating data acquisition. Hence, the first observed use of a hashtag in our data is possibly not the actual first use of that hashtag. To account for this uncertainty of initial hashtag usage time, we shift the initial tweet of each hashtag to the origin by an amount of time τ, such that $I(0) = 1$ in all cases, and add a variable I_{t-} to account for the existence of an informed population before the first hashtag detection. Therefore, (3.2) becomes a regression problem with four degrees of freedom: N, β, τ, and I_{t-}. The

least-squares objective function is defined as

$$\text{minimize} \sum_i |y(t_i) - I(t_i)|^2 \tag{3.3}$$

for all i data points of the given hashtag. Here, $y(t_i)$ are the observed data points, and $I(t)$ is given by

$$I(t) = \frac{Ne^{\beta(t-\tau)}}{N + (e^{\beta(t-\tau)} - 1)} - I_{t-}. \tag{3.4}$$

Since equation (3.4) requires a count of only the total population for $I(t)$ rather than the specific backbone network topology, the backbones are used to identify the subset of topic users whose collective hashtag adoption makes each $I(t)$ signal. The N, β, τ, and I_{t-} parameters are deduced from a non-linear least-squares regression of (3.4) on the set of $(t, I(t))$ points for each hashtag realization on a backbone network.

For each hashtag h that is tweeted on more than one topic backbone B, there exists a transmission rate parameter $\beta(h)$ and effective population size $N(h)$ for each of those backbones. In order to compare the $\beta(h)$ parameters for backbones of different effective population sizes, we must first normalize each $I(t)$ signal with respect to its best fit $N(h)$. By dividing both sides of equation (3.1) by N, one obtains

$$\frac{d\hat{I}}{d\hat{t}} = \hat{\beta}\hat{I}(1 - \hat{I}), \tag{3.5}$$

where $\hat{\beta} = \beta N$ and $\hat{I} = I/N$. It is also noted that substituting $\beta = \hat{\beta}/N$ into equation (3.2) leads to the normalized time scale $\hat{t} = t/N$. In this normalized setting, one interprets $\hat{\beta}$ as the number of interactions per unit of time (i.e., tweets among individuals that contain the hashtag of interest).

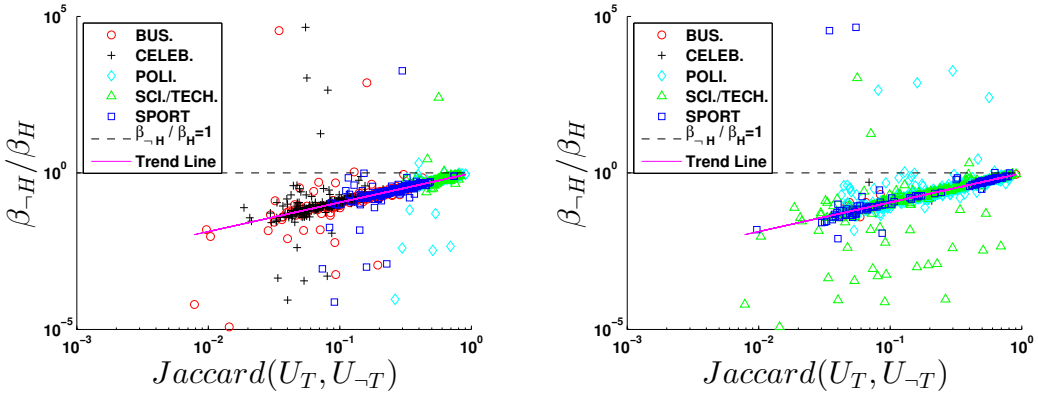


Figure 3.6: Relative transmission rate with respect to Jaccard similarity between two backbones on which a hashtag propagates in the SNAP dataset. The same data points are shown in both (a) and (b), but with different marking schemes, and each point in either plot represents a $(T, \neg T)$ pair. Color is added to improve marker differentiation. (a) Colors indicate the topic backbone on which a given hashtag h is propagating (i.e., colored by the $\neg T$ topic). (b) Colors indicate the true topic to which the given hashtag h belongs (i.e., colored by the T topic).

There are many hashtag users who are present on more than one topic backbone such that when one of these individuals uses a hashtag, that hashtag is observed to be simultaneously propagating on each topic backbone to which the user belongs. For example, suppose a Business related hashtag is used by an individual who is a member of the Business, Politics, and Sports topic backbones. The true topic (T) of this particular hashtag is Business, and a not true topic ($\neg T$) is either Politics or Sports. In this case, there will be two $(T, \neg T)$ pairs: (Business,Politics) and (Business,Sports).

We denote the transmission rate of the hashtag on its actual topic backbone $\hat{\beta}_T(h)$ and the hashtag transmission rate on an off-topic backbone as $\hat{\beta}_{\neg T}(h)$. For each hashtag, we also denote the Jaccard similarity between the subset of those hashtag users on the back-

bones of a $(T, \neg T)$ pair as Jaccard$(U_T(h), U_{\neg T}(h))$, where $U_T(h) := \{u \in B_T \mid \forall u \in (U, h)\}$ and $U_{\neg T}(h) := \{u \in B_{\neg T} \mid \forall u \in (U, h)\}$. Recall that B represents the topic backbone, and should not be confused with β, which represents the transmission rate of (3.1).

Figures 3.6a and 3.6b show the data comparing $\hat{\beta}_T(h)$ relative to each $\hat{\beta}_{\neg T}(h)$ in the vertical dimension, and the Jaccard similarity of the respective users of h in the corresponding T and $\neg T$ backbones, in the horizontal dimension. Overall, we see that, on average, each hashtag propagates fastest on its own topic network since an overwhelming majority of the data points lie below the $\hat{\beta}_{\neg T}(h)/\hat{\beta}_T(h) = 1$ line.

Figure 3.6a demonstrates that the relative rates of propagation tend to increase as the topic backbones increasingly overlap. This is particularly evident for the Business, Celebrity, and Sports topic backbones. The collection of Sci./Tech points below the trend line of Figure 3.6b indicates that these hashtags have transmission rates on off-topic backbones $\hat{\beta}_{\neg T}(h)$ that are much less than their true topic backbone $\hat{\beta}_T(h)$. The corresponding points in Figure 3.6b indicate which off-topic backbone yields the transmission rate $\hat{\beta}_{\neg T}(h)$.

Outliers in Figures 3.6a and 3.6b are an artifact of the SI-model not being an appropriate underlying model for their data, but are included in the results because either the T or $\neg T$ backbones for the associated hashtag proved to have SI-type behavior. The outliers, however, have little effect on the trend line shown in Figures 3.6a and 3.6b, since the trend line has an average point-wise residual of 0.15 on the log-log scale shown.

Chapter 4

Data Parameter and Uncertainty Estimation

Thus far, we have been able to track hashtags on the Twitter network, and have shown that the characteristic logistic adoption behavior of the population is captured on the topic backbone structures using standard least-squares regression techniques. However, least squares regression alone does not sufficiently quantify the variance about the mean-field signal of the least-squares solution, and quantifying this variance is essential for being able to draw conclusions about what effects can be attributed to the network structure.

In this section, we describe how statistical sampling methods, namely the delete-d jackknife sampling method, can be used to estimate the variance about the mean-field solution to the least-squares regression problem. We shall also go on to describe how data can be assimilated in real-time using an adaptive jackknife estimation strategy. The jackknifed estimation of the data signal synthetically estimates the underlying data distribution, which can be described by the initial parameter distributions of the SI model.

The agent-based network simulations discussed in Chapter 2 can then be run with the same initial distribution so that the homogeneous solution to the agent-based model agrees with the ensemble mean-field solution of the jackknifed data. From here we can run a paired statistical hypothesis test to determine how likely the agent-based model explains the data. We find that the agent-based network model does not adequately explain the data, which suggests that a more precise, yet efficient, user model is needed.

4.1 Stochastic model estimation

Let us consider a continuous nonlinear model that contains model uncertainty in the form of a stochastic forcing term, and is measured at discrete instances of time t_k:

$$dx = f(t, x)dt + \sqrt{Q}dw, \tag{4.1a}$$

$$y(t_k) = h(x(t_k)) + \sqrt{R}N(0, 1), \tag{4.1b}$$

where $x \in \mathbb{R}^n$ represents the state of the system, $f(t, x) : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is the deterministic evolution of the states, $h(x) : \mathbb{R}^n \to \mathbb{R}^m$ is a function that maps x to the discrete-time measured output $y \in \mathbb{R}^m$, dw describes a vector Wiener process with mean zero and unit variance, and $N(0, 1)$ represents a normally distributed random variable with zero mean and unit variance. It is also noted that the covariance matrices Q and R are positive semi-definite symmetric matrices, and their square roots exist and can be computed using a singular value decomposition [60].

Since its discovery, the Kalman filter, in both its linear and nonlinear forms, has been an effective model-based noise filter that relies on an assumed known deterministic model with additive noise [61]. However, when either parameters of the model or noise variances are unknown, which is common in tasks where model identification and state estimation must occur simultaneously, the Kalman filter is likely to diverge [62, 63].

To prevent divergence, various tuning procedures exist for finding the best estimates of process and measurement noise for given a Kalman filter [64–68]. Kalman filter tuning typically involves minimizing the measurement error over iterations of the Kalman filter, with the process and measurement covariances as the free variables. For nonlinear systems, this type of tuning procedure requires that at each optimization iteration, the gradient of a complete time sequence of Kalman filter iterations is taken with respect to all of the free variables. Hence, these methods are computationally costly, and are susceptible to converging to suboptimal local minima of their objective functions.

Adaptive algorithms have also been developed to allow the Kalman filter to converge on the correct noise values in an online manner [69–79]. Much effort has been given to developing adaptive methods for nonlinear systems because online computation is in the spirit of the Kalman filter. Adaptive methods for linear systems have seen much success over the years [77], but their formulation is limited to the linear case and does not extend to nonlinear systems in general. For nonlinear systems, adaptive strategies have been implemented for the main variants of the nonlinear Kalman filter: the extended Kalman filter (EKF) [74], and the unscented Kalman filter (UKF) [69, 72, 73]. However, for

the adaptive EKF and UKF methods, convergence performance has yet to be rigorously generalized, and is sensitive to the initial estimates of the unknown parameters.

To overcome these challenges associated with implementing adaptive nonlinear Kalman filters, we propose that the unknown state and parameter distributions of the given model can be estimated by an ensemble of least-squares regression (LSQ) estimates on the known data. Jackknife sampling methods [80–82] can be used to generate the ensemble of LSQ estimates [83], and this ensemble generation procedure can then be made adaptive (in a Markov-Chain sense) by taking advantage of how jackknife sampling assimilates newly acquired data into the model. The formula for a statistical Kalman filter can then be used to infer the unknown process uncertainty and measurement noise covariance matrices from ensemble estimates at each step. After the unknown quantities of the stochastic model have converged, the adaptive procedure can be stopped, and a standard nonlinear Kalman filter can be implemented to take over the state estimation process.

Although our approach is supported by the theory behind ensemble Kalman filtering (EnKF) [84, 85], our adaptive method of assimilating the data is original, as well as our application of jackknife sampling to generate ensemble members. Particle filters and the EnKF both make assumptions on the sampling distribution of states, and typically rely on Markov-Chain Monte Carlo (MCMC) simulation to generate ensemble members and deduce ensemble statistics of the states. We show that by using LSQ estimation in conjunction with jackknife sampling of the known data, a sampling distribution and ensemble statistics can be acquired without making any assumptions of the sampling

distribution nor having to run a high number of MCMC simulations. Furthermore, we describe how our adaptive method can be implemented in a parallel setting, and with a fixed number of computations at each update step.

Therefore, the aim of this manuscript is to implement the techniques of jackknife variance estimators as they apply to least squares estimators, to construct an adaptive, nonparametric, and computationally efficient statistical nonlinear filter. To motivate our jackknife sampling LSQ approach for generating ensemble statistics, we shall present an overview of the derivation of a statistical Kalman filter in Section 4.2. In Section 4.3 we present a description of jackknife sampling methods, an adaptive jackknife sampling approach to assimilating new data, and how jackknife sampling can be used with LSQ estimation. We combine the results of Sections 4.2 and 4.3 to construct a procedure in Section 4.4 for estimating the process and measurement noise of the model (4.1). We present an example application in Section 4.5 to demonstrate the efficacy of our adaptive jackknife filter, and we summarize our conclusions and directions for future work in Section 4.6.

4.2 Ensemble Kalman Filtering

The ultimate objective of this manuscript is to develop a procedure to optimally estimate the states of a noisy nonlinear state-space model, when only a model structure and some observed measurements are provided. Since only a model structure is assumed, we shall attempt to estimate model parameters while simultaneously estimating model

states. Optimal state estimation is often performed using a Kalman filter, which has many linear and nonlinear variants. For reasons that will be discussed later, we shall focus our attention on the EnKF and its formulation [61, 84, 85]. Here, we will describe the EnKF in order to motivate our approach for estimating the unknown model parameters in the next section.

4.2.1 Model Uncertainty Propagation

Let us consider a general nonlinear model that contains model uncertainty in the form of a stochastic forcing term

$$dx = f(t, x)dt + g(x)dq, \tag{4.2}$$

where x represents the state of the system, $f(t, x)$ gives the deterministic evolution of the states, $g(x)$ is a function that may depend on the states, and $dq = \sqrt{Q}dw$ describes a vector Wiener process with mean zero and covariance matrix $Q\delta(t)$. It is noted as a technical detail that since $g(x)$ is not an explicit function of dq, the Ito interpretation is used [86], and $\int_{t_{k-1}}^{t_k} dw = \sqrt{t_k - t_{k-1}}N(0, 1)$. Thus, one can integrate (4.2) from t_{k-1} to t_k to obtain the distribution of $x(t_k)$ when the distribution $x(t_{k-1})$ is known. The probability distribution of $x(t_k)$ for a given initial point $x(t_{k-1})$ is

$$x(t_k) = F(t_k, x(t_{k-1})) + g(x(t_{k-1}))\sqrt{Q\Delta t_k}N(0, 1), \tag{4.3}$$

where $\Delta t_k = (t_k - t_{k-1})$, and $F(t_k, x(t_{k-1}))$ is the evolution operator that deterministically maps x from time t_{k-1} to t_k according to the $dx = f(t, x)dt$ part of (4.2).

When $g(x)dq$ is normally distributed and forms a Markov process, it is shown in [84] that it is possible to derive the Fokker-Planck equation to describe the time evolution of the probability density function $p(x,t)$ of the model state:

$$\frac{\partial p(x,t)}{\partial t} + \sum_i \frac{\partial (f_i(t,x)p(t,x))}{\partial x_i} = \frac{1}{2}\sum_{i,j} \frac{\partial^2 (gQg^T)_{ij}}{\partial x_i \partial x_j}, \tag{4.4}$$

where $f_i(t,x)$ is the i^{th} component of $f(t,x)$, and gQg^T is the covariance matrix for the model errors at time t.

The EnKF, as discussed in [84] and [85], applies a Markov Chain Monte Carlo Method (MCMC) to solve (4.4). The probability density $p(x,t)$ is represented by an ensemble of N model states $x^{(i)}$ for $i \in \{1,\ldots,N\}$, and the ensemble prediction, by integrating model states forward according to (4.3), is equivalent to using a MCMC method to solve (4.4). Hence, there is no need to find an explicit form for the solution $p(x,t)$ of (4.4) because $p(x,t)$ can be sufficiently described by its ensemble statistics.

Since we assume no prior knowledge of the function $g(x)$, we shall simplify matters and take $g(x) = I_{n\times n}$ so that all of the model uncertainty is spatially invariant and entirely attributed to the process noise. Furthermore, we shall assume discrete measurements y_k at times t_k, which have their own uncertainty that we shall assume to be normally distributed. For the remainder of the manuscript we shall assume the continuous-discrete stochastic model defined by (4.1).

4.2.2 Statistical Derivation of a Kalman Filter

To help explain how the Kalman filter is implemented from an ensemble of nonlinear system realizations, it is instructive to first provide a statistical derivation of the Kalman filter. The following statistical derivation for a general Kalman filter closely follows [61]. We define the following variables at the discrete time instance t_k of the latest measurement $y(t_k)$:

- $x(t_k)$ = true state value,
- $\widehat{x}^-(t_k)$ = state estimate prior to measurement,
- $\widehat{x}^+(t_k)$ = posterior state estimate,
- $P^-(t_k) = E\left[(x(t_k) - \widehat{x}^-(t_k))([\cdots])^T\right]$,
- $P^+(t_k) = E\left[(x(t_k) - \widehat{x}^+(t_k))([\cdots])^T\right]$,

where the $[\cdots]$ is shorthand notation for the term immediately to the left of it, so that the covariance matrices are written as $E\left[(z)([\cdots])^T\right] = E\left[(z)(z)^T\right]$.

Suppose, for computational performance reasons, we want a state estimator that linearly updates the ensemble mean of its state estimate $\overline{\widehat{x}}(t_k)$ based on the latest measurement $y(t_k)$ according to the rule

$$\overline{\widehat{x}^+}(t_k) = K(t_k)y(t_k) + b(t_k), \tag{4.5}$$

where $K(t_k)$ and $b(t_k)$ are a yet to be determined matrix and vector, respectively. For notational convenience, we shall momentarily omit any explicit dependence on t_k because

all of the variables are understood to be implicitly evaluated at the same time instance t_k.

Since we want an unbiased state estimate (i.e., $\overline{\widehat{x}^+} = \overline{x}$), we can see by taking the mean of (4.5) that

$$\overline{\widehat{x}^+} = K\overline{y} + b, \tag{4.6}$$

which gives the constraint that

$$b = \overline{x} - K\overline{y}, \tag{4.7}$$

which ensures unbiasedness of the estimate $\widehat{x}^+$ regardless of K.

To find the gain matrix K, we shall solve for the K that minimizes the expression for the trace of P_x^+. In general, a covariance is defined as

$$\begin{aligned} P_z &= E\left[(z-\overline{z})(z-\overline{z})^T\right], \\ &= E\left[zz^T\right] - \overline{z}\overline{z}^T, \end{aligned} \tag{4.8}$$

for any random vector z. Let us now set $z = x - \widehat{x}^+$. Because of the unbiasedness of the estimate of $\widehat{x}^+$ as asserted by (4.7), it is noted that $\overline{z} = 0$, and

$$\begin{aligned} P_x^+ &= E[zz^T] \\ &= P_z + \overline{z}\overline{z}^T \\ &= P_z. \end{aligned} \tag{4.9}$$

By directly calculating P_z for $z = x - \widehat{x}^+$, one obtains

$$\begin{aligned} P_x^+ &= P_z \\ &= E\left\{[x - \widehat{x}^+ - E(x - \widehat{x}^+)][\cdots]^T\right\} \\ &= E\left\{[x - (Ky + b) - \overline{x} + (K\overline{y} + b)][\cdots]^T\right\} \\ &= E\left\{[(x - \overline{x}) - K(y - \overline{y})][\cdots]^T\right\} \\ &= P_x - KP_{yx} - P_{xy}K^T + KP_yK^T, \end{aligned} \tag{4.10}$$

where P_{ab} denotes the cross-covariance of random variables a and b. Using the fact that covariance matrices are symmetric (i.e., $P_{xy} = P_{yx}^T$), then the trace of (4.10) becomes

$$\begin{aligned} \mathrm{Tr}(P_x^+) &= \mathrm{Tr}(P_x - KP_{yx} - P_{xy}K^T + KP_yK^T) \\ &= \mathrm{Tr}\left[(K - P_{xy}P_y^{-1})P_y(K - P_{xy}P_y^{-1})^T\right] \\ &+ \mathrm{Tr}\left[P_x - P_{xy}P_y^{-1}P_{xy}^T\right]. \end{aligned} \tag{4.11}$$

Since covariance matrices are positive semi-definite, the first term of (4.11) is non-negative, and is identically equal to the zero matrix when $K = P_{xy}P_y^{-1}$. The second term of (4.11) does not depend on K. Therefore, the posterior covariance estimate P_x^+ is minimized when

$$K = P_{xy}P_y^{-1}. \tag{4.12}$$

If the prior estimate $\widehat{x}^-$ is also unbiased so that $\overline{x} = \overline{\widehat{x}^-}$, then $P_x = P_x^-$. By making this assumption and substituting equations (4.12) and (4.7) into (4.10), we have the

update equations

$$\overline{\widehat{x}^{+}} = \overline{\widehat{x}^{-}} + K\left(y - \overline{y}\right), \tag{4.13}$$

$$P_x^{+} = P_x^{-} - KP_{xy}^{T}. \tag{4.14}$$

We remark that equations (4.7) and (4.13) have the Markov property, and are only true when the Markov property is true for the equations that determine each of these elements. In the next section we will discuss how one can use the ensemble output statistics to appropriately estimate $\overline{y}$, and prevent measurement bias from affecting the state estimate in (4.13).

4.2.3 Ensemble estimation of P_x and P_y

When integrating an ensemble of points forward in time according to (4.3), the state covariance matrix P_x^{-} depends on the distribution of those deterministic points and the stochastic forcing term. For notational convenience, let us denote $\widehat{x}^{-} = F(t_k, x(t_{k-1}))$. Since the ensemble mean is unbiased so that $\overline{\widehat{x}^{-}} = \overline{x}$, then an approximation for the prior ensemble covariance $P_x^{-}(t_k)$ becomes

$$\begin{aligned}
P_x^{-} &= E\left[(x - \overline{\widehat{x}^{-}})(\cdots)^T\right] \\
&= E\left[(\widehat{x}^{-} - \overline{\widehat{x}^{-}} + \sqrt{Q\Delta t_k}N(0,1))(\cdots)^T\right] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[(\widehat{x}_{(i)}^{-} - \overline{\widehat{x}^{-}})(\cdots)^T\right] + Q\Delta t_k \\
&= \widehat{P_x}^{-} + Q\Delta t_k,
\end{aligned} \tag{4.15}$$

where $\widehat{P_x^-}$ is the sample ensemble covariance of the state prior distribution, and

$$\overline{\widehat{x^-}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{x}_{(i)}^-$$

for the collection of N ensemble members $\widehat{x}_{(i)}^-$.

To find the measurement covariance P_y and cross-covariance P_{xy}, the process noise can be made an explicit term by taking a series expansion of $h(x(t_k))$ about $\widehat{x}^-$ at time t_k:

$$h(x(t_k)) = h(\widehat{x}^- + \sqrt{Q\Delta t_k}N(0,1)) \tag{4.16}$$

$$= h(\widehat{x}^-) + Dh_{\widehat{x}^-}\sqrt{Q\Delta t_k}N(0,1) + \sum_{n=2}^{\infty}\frac{1}{n!}D^n h_{\widehat{x}^-}(\sqrt{Q\Delta t_k}N(0,1))^n, \tag{4.17}$$

where $D^n h_{x^{(i)}}$ represents the n^{th} vector derivative of h about the point $x_{(i)}(t_k)$. From this we obtain

$$\begin{aligned}
P_y &= E\left[(y-\overline{y})(\cdots)^T\right] \\
&= E\left[(h(x(t_k)) - \overline{h(x(t_k))} + \sqrt{R}N(0,1))(\cdots)^T\right] \\
&= E\Big[(h(\widehat{x}^-) + Dh_{\widehat{x}^-}\sqrt{Q\Delta t_k}N(0,1) - \overline{h(\widehat{x}^-)} \\
&+ \sqrt{R}N(0,1))(\cdots)^T\Big] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[(h(\widehat{x}_{(i)}^-) - \overline{h(\widehat{x}_{(i)}^-)})(\cdots)^T\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N} Dh_{\widehat{x}_{(i)}^-} Q\Delta t_k Dh_{\widehat{x}_{(i)}^-}^T + R \\
&= \widehat{P}_y + \widehat{Q}_y + R,
\end{aligned} \tag{4.18}$$

where $\widehat{P}_y$ is the sample ensemble covariance of the measurements and $\widehat{Q}_y$ comes from the stochastic forcing term. Similarly, one finds the cross covariance to be

$$\begin{aligned} P_{xy} &= E\left[(x-\overline{x})(y-\overline{y})^T\right] \\ &= E\left[(\widehat{x}^- - \overline{x})(h(\widehat{x}^-) - \overline{h(\widehat{x}^-)})^T\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left[(\widehat{x}^-_{(i)} - \overline{\widehat{x}^-_{(i)}})(h(\widehat{x}^-_{(i)}) - \overline{h(\widehat{x}^-_{(i)})})^T\right] \\ &= \widehat{P}_{xy}, \end{aligned} \tag{4.19}$$

where $\widehat{P}_{xy}$ is the sample ensemble cross-covariance and all additive noise terms vanish because they are mutually uncorrelated.

By substituting equations (4.15), (4.18), and (4.19) into equations (4.12) and (4.14), one obtains

$$K = \widehat{P}_{xy}(\widehat{P}_y + \widehat{Q}_y + R)^{-1} \tag{4.20}$$

$$\overline{\widehat{x}^+} = \overline{\widehat{x}^-} + K\,(y - \overline{y}) \tag{4.21}$$

$$P_x^+ = \widehat{P}_x^- + Q\Delta t_k - K(\widehat{P}_y + \widehat{Q}_y + R)K^T. \tag{4.22}$$

However, to implement this nonlinear statistical filter, we need to have quantities for Q and R, which we propose can be estimated directly from the data, and without making any assumptions on their sampling distribution. We did make assumptions that the process and measurement noise terms are Gaussian, which we will find in the next section, is actually consistent with a least squares parameter estimation strategy.

4.3 Ensemble generation and adaptive update

In order to implement a statistical Kalman filter, we need to obtain ensemble estimates of the state and output distributions. To do this, we can take a statistical sample of those distributions via *jackknife sampling* [80–82], which has been shown to be a robust and computationally efficient way of estimating the sample distribution of a given population. By mapping the data to the state-space via LSQ estimation, we shall jackknife sample the known data in order to obtain the underlying sample distribution of the states and model parameters. The points that define the sample distribution of the states and model parameters are then treated as ensemble members for the statistical Kalman filter. We shall first explain jackknife sampling, its consistency properties and an adaptive update rule, and then apply jackknife sampling to LSQ estimation.

4.3.1 Jackknife Sampling

Suppose we are given a sequence of n data measurements $D_n = \{Y_1, \ldots, Y_n\}$, where $Y_i = (y_i, t_i)$ is defined for an observed response vector y_i from a known input sequence of t_i values. For the moment, let us fix the number of available data points n and choose some fixed positive integer d. We shall describe the *delete-d jackknife* estimator [81, 82], which estimates the sample distribution of parameters by aggregating the least squares estimates on randomly chosen subsets of $r = n - d$ data points. Let S_r be the collection of subsets of $\{1, \ldots, n\}$ that have size r. For $s = \{i_1, \ldots, i_r\} \in S_r$, let $\widehat{\theta}_s = \widehat{\theta}\,(Y_{i_1}, \ldots, Y_{i_r})$.

The delete-d jackknife estimator of $\text{var}(\theta_n)$ is defined as

$$v_n = \frac{r}{dN} \sum_{s \in S_r} \left(\widehat{\theta}_s - \theta_n\right) \left(\widehat{\theta}_s - \theta_n\right)^T, \tag{4.23}$$

where $N = \binom{n}{d}$, and θ_n is the parameter estimate that explains all of the available n data points. For a finite set of measurements, we can approximate θ_n by the arithmetic average of subsample means, which we call the jackknife estimate $\widehat{\theta}_n$, and define

$$\widetilde{v}_n = \frac{r}{dN} \sum_{s \in S_r} \left(\widehat{\theta}_s - \widehat{\theta}_n\right) \left(\widehat{\theta}_s - \widehat{\theta}_n\right)^T \tag{4.24}$$

with

$$\widehat{\theta}_n = \frac{1}{N} \sum_{s \in S_r} \widehat{\theta}_s.$$

When N is very large, the number of computations can be reduced by implementing techniques from survey sampling. For instance, take a simple random sample (without replacement) of size m from S_r (i.e., $S_m \subset S_r$). Compute $\widehat{\theta}_s$ for $s \in S_m$, and use

$$v_n^s = \frac{r}{dm} \sum_{s \in S_m} \left(\widehat{\theta}_s - \theta_n\right) \left(\widehat{\theta}_s - \theta_n\right)^T \tag{4.25}$$

and

$$\widetilde{v}_n^s = \frac{r}{dm} \sum_{s \in S_m} \left(\widehat{\theta}_s - \widehat{\theta}_n\right) \left(\widehat{\theta}_s - \widehat{\theta}_n\right)^T \tag{4.26}$$

with

$$\widehat{\theta}_n = \frac{1}{m} \sum_{s \in S_m} \widehat{\theta}_s.$$

to approximate v_n and $\widetilde{v}_n$, respectively. These approximations are called the *jackknife-sampling variance estimators* (JSVE's) [81, 82], and m is the second-stage sample size.

It is also noted that the pre-factor terms $r/(dN)$ and $r/(dm)$ are explained in [81, 82], and mitigate the bias associated with estimating the variance from a finite sample.

In [87], it was shown that

- ([87] Theorem 1) $\text{var}(v_n) = o\left(n^{-2}\right)$,
- ([87] Theorem 2) $0 \leq \text{var}(v_n^s) - \text{var}(v_n) = O(m^{-1}\tau_n)$, for $\tau_n = E\left[(\theta_n - \theta)^4\right]$.

We remark that $\text{var}(v_n^s)$, $\text{var}(v_n)$, and $E\left[(\theta_n - \theta)^4\right]$ are well defined for jointly distributed random variables [88], and are only needed here to prove asymptotic consistency of jackknife sampled distributions.

The authors of [87] also show that choosing $m = n^\delta$ for some $\delta \geq 1$ is sufficient and has the same number of computations as the delete-1 jackknife estimator. If m is much smaller than N, sampling with replacement for the second-stage sample will produce almost the same estimator as sampling without replacement, which further simplifies the sampling procedure and is nearly identical to bootstrap sampling. It is also important to note that these results do not necessarily rely on $m^{-1}\sum_{s\in S_m}\theta_n \to \theta$ as $m \to \infty$, which is a convergence result that we will further discuss next.

4.3.2 Adaptive Jackknife Variance Estimator

Although the estimates are conditioned on past data, we see that the ensemble jackknife estimates abide by the Markov property in the sense that they only rely on the previous ensemble measurement and the current ensemble measurement. When tracking

only the mean and variance of the distribution, all of the previous ensemble members may be forgotten, as their statistics are sufficiently captured by the mean and variance.

Suppose another measurement is collected so that there are now a total of $n+1$ data points, and for computational reasons we want the values of r and m to remain the same as before. When constructing the basic form of our adaptive equations, it is important to define the mean and variance of the linear combination of two uncorrelated random variables X_1 and X_2. For $\mu_1 = E[X_1]$, $\mu_2 = E[X_2]$, $v_1 = \text{var}(X_1)$, $v_2 = \text{var}(X_2)$, and two constants $a_1, a_2 \in \mathbf{R}$ such that

$$X_3 = a_1 X_1 + a_2 X_2,$$

then

$$E[X_3] = a_1\mu_1 + a_2\mu_2, \tag{4.27}$$

$$\text{var}(X_3) = a_1^2 v_1 + a_2^2 v_2. \tag{4.28}$$

In this context, each jackknife estimate $\widehat{\theta}_n$ can be viewed as a combination of jackknife estimates $\widehat{\theta}_{n\in s}$ that include the n^{th} data point, and those that do not $\widehat{\theta}_{n\notin s}$:

$$\widehat{\theta}_n = a_1\widehat{\theta}_{n\in s} + a_2\widehat{\theta}_{n\notin s}, \tag{4.29}$$

where $a_1 + a_2 = 1$. It is also assumed that $\widehat{\theta}_{n\in s}$ and $\widehat{\theta}_{n\notin s}$ are uncorrelated, which is intuitively justified by the fact that the noise contributing to the n^{th} data point is uncorrelated with the noise contributing to any of the previous $n-1$ data points.

The values a_1 and a_2 in (4.29) represent the relative likelihoods of occurrence for the two types of jackknife estimates $\widehat{\theta}_{n\in s}$ and $\widehat{\theta}_{n\notin s}$, respectively. If we temporarily remove the

n^{th} data point from the data set, we see that there are $\binom{n-1}{r}$ possible unique jackknife estimates $\widehat{\theta}_{n\notin s}$ that can be obtained from r data points. Moreover, it becomes apparent that $\widehat{\theta}_{n\notin s} = \widehat{\theta}_{n-1}$. Since there are $\binom{n}{r}$ total possible unique jackknife estimates of $\widehat{\theta}_n$, the likelihood of reselecting an estimate $\widehat{\theta}_n$ is $\binom{n-1}{r}\binom{n}{r}^{-1} = 1 - r/n$. Hence, one obtains

$$a_1 = r/n \text{ and } a_2 = 1 - r/n. \tag{4.30}$$

By substituting equations (4.29) and (4.30) into (4.27), and observing that $\widehat{\theta}_{n\notin s} = \widehat{\theta}_{n-1}$, the adaptive jackknife sample mean estimator is defined to be

$$\overline{\widehat{\theta}_n} = \frac{r}{n}\overline{\widehat{\theta}_{n\in s}} + \left(1 - \frac{r}{n}\right)\overline{\widehat{\theta}_{n-1}}, \tag{4.31}$$

where

$$\overline{\widehat{\theta}_{n\in s}} = \frac{1}{m}\sum_{s\in S_m^+}\widehat{\theta}_s,$$

and

$$S_m^+ = \{s \in S_m | n+1 \in s = \{i_1, \ldots, i_m\}\}.$$

Similarly, the jackknife sample variance update is obtained by substituting equations (4.29) and (4.30) into (4.28). By again observing that $\widehat{\theta}_{n\notin s} = \widehat{\theta}_{n-1}$, one gets

$$\tilde{v}_n^s = \left(\frac{r}{n}\right)^2 \tilde{v}_{n\in s}^s + \left(1 - \frac{r}{n}\right)^2 \tilde{v}_{n-1}^s, \tag{4.32}$$

where

$$\tilde{v}_{n\in s}^s = \frac{r}{(d+1)m}\sum_{s\in S_m^+}\left(\widehat{\theta}_s - \overline{\widehat{\theta}_n}\right)\left(\widehat{\theta}_s - \overline{\widehat{\theta}_n}\right)^T.$$

Equation (4.31) inherits the convergence properties of its respective constituent terms $\widehat{\theta}_{n\in s}$ and $\widehat{\theta}_{n\notin s}$, because each of those constituent terms have identical convergence properties and (4.31) is a convex combination of its constituent terms. The same reasoning

about convergence applies to (4.32) and its constituent terms $\tilde{v}^s_{n\in s}$ and $\tilde{v}^s_{n\notin s}$, as well. Furthermore, since we are effectively keeping track of a running average of second-stage m samples, the total number of second-stage samples acquired at measurement number $n = n_0 + k$ is $m_n = m_0 + km$, where m_0 is the number of second-samples used to estimate the first n_0 measurements. By choosing $m_0 = n_0$, then the condition $m_n = n^\delta$ for some $\delta \geq 1$ is satisfied, and the variance estimate of v_n has the same accuracy as the delete-1 jackknife, but for a fixed number of computations at each increment of n.

4.3.3 Least Squares Parameter Estimator

The previous sections established general results for the convergence in sample-variance for a parameter estimate without any mention of the parameter estimator. Since we want to make no assumptions about the parameter's prior distribution, we shall choose the well known LSQ estimator. Fortuitously, the LSQ estimator naturally produces a normally distributed parameter estimate [83], which is consistent with the assumed uncertainty terms in the stochastic model (4.1a) and (4.1b).

Suppose we have, again, a sequence of n data measurements $D_n = \{Y_1, \ldots, Y_n\}$, where $Y_i = (y_i, t_i)$, as defined earlier. Adopting much of the notation from [83], we consider a general nonlinear model to describe an observed sequence of data

$$y_i = H(t_i, \theta) + \sigma e_i, \qquad i = 1, \ldots, n, \tag{4.33}$$

where θ is a vector of unknown constant parameters, $H(t, \theta)$ is a nonlinear function in θ, the e_i's are independent and identically distributed (i.i.d.) unobservable random variables

with mean zero and variance one, and σ is the unknown error standard deviation. It is also noted that the error terms define the measurement residuals $r_i = (y_i - H(t_i, \theta)) = \sigma e_i$.

A LSQ parameter estimator finds an estimate $\widehat{\theta}_n$ of the parameters that minimizes the mean squared error (MSE) for a model over all available data points

$$\widehat{\theta}_n = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - H(t_i, \theta))^2, \tag{4.34}$$

which effectively minimizes σ in the model (4.33). In relation to the SDE model (4.2), one finds that $H(t, \theta) = h(F(t, x(T)))$ when $\theta = x(T)$ for some fixed point in time T.

We remark that the solution to (4.34) also minimizes the sample variance of the $\widehat{\theta}_n$ estimate's residuals $\operatorname{var}(\widehat{r}_n)$. When using all of the data points, the solution to (4.34) is only one point estimate of the parameters. With only one point estimate of the parameters $\widehat{\theta}_n$, there is no knowledge about how sensitive the parameters are to the data, or equivalently, what the variance estimate is of the parameters (i.e. $\operatorname{var}(\widehat{\theta}_n)$) that produced the given realization of the data. Jackknife variance estimation, such as the JSVE, provides a way of aggregating parameter estimates without making any prior assumptions about the distribution of $\widehat{\theta}$ (i.e., JSVEs are *nonparametric* estimators).

From the given data realization D_n, we can implement a delete-d jackknife sampling of D_n to generate a sample distribution of D, which directly gives us a sample distribution of θ by running the LSQ estimator on each jackknife sample of D_n. This approach is rigorously studied in [83] (and references therein), which specifically describes the asymptotic consistency properties of the LSQ estimator and its jackknife variance estimator in nonlinear models. For the jackknife estimate $\widehat{\theta}_n$ of θ_n, it was found in [83] that consis-

tency and asymptotic normality of $\widehat{\theta}_n$ can be established, as well as the consistency of the jackknife variance estimator of the asymptotic covariance matrix of $\widehat{\theta}_n$. The results are summarized here for the delete-1 jackknife, as originally presented in [83], and can easily be extended to the delete-d case using the results of the previous sections.

- ([83] Theorems 1 and 2) For a LSQ estimator θ_n conditioned on n data points, then $\theta_n \to \theta$ almost surely (a.s.), and the distribution of a sequence of consistent LSQ estimators θ_n is asymptotically normally distributed.

- ([83] Lemma 3) Let $\widehat{\theta}_s$, for $i = \{1, \ldots, n\}$, be the collection of delete-1 jackknife samples of the LSQ estimates of θ_n. Then

$$\max_{i \leq n} \left\| \widehat{\theta}_{ni} - \theta \right\| \to 0 \qquad \text{a.s.} \tag{4.35}$$

- ([83] Theorem 4) The jackknife variance estimator is consistent, by proving that $n(\tilde{v}_n - v_n) \to 0$ a.s.

Therefore, a jackknifed sampling of least squares estimates allows us to estimate a prior distribution of parameters for a nonlinear model without having to implement MCMC methods. An added benefit of the LSQ jackknife sampling procedure is that the estimated parameter distribution will asymptotically be normally distributed. Ensuring that the distributions are normal is essential to the performance of the EnKF, since the EnKF only uses the first two moments of the ensemble distribution. Furthermore,

the adaptive scheme in the previous section provides a computationally efficient way of assimilating new data into the statistical model.

4.4 Posterior estimation via ensemble filtering

In previous sections, we saw how to use ensemble filtering to construct posterior estimates of a distribution's mean and covariance without having to implement MCMC methods. However, the ensemble filtering requires knowledge of the process noise, measurement noise, and the mean and covariance of the prior distribution. When these prior quantities are known, ensemble filtering can be implemented to further reduce the computational cost of assimilating new data. Without prior knowledge of model parameters or model noise distributions, we propose that one can implement jackknife estimation methods to initialize the stochastic model such that ensemble filtering can take over the posterior parameter and state estimation process once it produces posterior estimates that agree with that of the adaptive jackknife method.

4.4.1 Estimating R from Cross-Validation

When implementing the jackknife LSQ estimator, the sampling distribution for θ produces an output distribution for y. However, the measurements are subject to uncertainty, as accounted for in (4.2), and this uncertainty can be measured as being attributed to the additional *out-of-sample* error. Cross-validation (CV) is a statistical learning technique typically used to evaluate a model by describing its out-of-sample statistics. Typical

CV methods involve training a model on a subset S_m of the available data, and then validating (testing) the model on the complement of S_m, which we denote as S_m^c.

The delete-d jackknife variance estimator already removes d data points from the available data before each step of the parameter estimation, which naturally allows us to use those d data points to acquire out-of-sample residual statistics that are indicative of the errors we would see for a future measurement. Furthermore, we can use the delete-d jackknife methodology to obtain jackknife estimates of the residual statistics, except the validation set uses a delete-r jackknife estimate.

For a given jackknife parameter estimate $\widehat{\theta}_s$ such that $s \in S_m$, a residual $\widehat{r}_j$ is defined for some $j \in S_m^c$ as

$$\widehat{r}_j = y_j - H(t_j, \widehat{\theta}_s), \tag{4.36}$$

and the residual statistics defined for a set of μ indices $\{j_1, \ldots, j_\mu\} \in S_m^c$, for which $\mu \leq d$, are

$$\overline{\widehat{r}_s} = \frac{1}{\mu} \sum_{j \in S_m^c} \widehat{r}_j,$$

$$\widehat{\sigma}_s^2 = MSE(\widehat{\theta}_s) = \frac{d}{r\mu} \sum_{j \in S_m^c} \widehat{r}_j \widehat{r}_j^T,$$

where $\widehat{\sigma}_s^2$ estimates the out-of-sample variance of $\widehat{\theta}_s$.

For each $\widehat{\theta}_s$ estimate, there exists a corresponding jackknife sample distribution of out-of-sample residual values $\widehat{r}_j$. The jackknife mean of $\widehat{r}$ is the measurement bias, and the jackknife variance estimate $\widehat{\sigma}_s^2$ captures the uncertainty attributed to both $\widehat{\theta}_n$ and the measurement noise $\sqrt{R}N(0,1)$. Since we obtain m estimates of $\widehat{\theta}_s$, we also

obtain m sample distributions of the out-of-sample residuals, and the expected residual distribution is described by the arithmetic mean of the m residual distributions (i.e., each residual distribution has equal probability of being the correct residual distribution). The expected jackknife residual statistics are

$$\overline{\widehat{r}_n} = \frac{1}{m} \sum_{s \in S_m^c} \widehat{r}_s, \tag{4.37}$$

$$\widehat{\sigma}_n^2 = \frac{1}{m^2} \sum_{s \in S_m^c} \widehat{\sigma}_s^2. \tag{4.38}$$

The adaptive rule outlined in Section 4.3 can also be applied to obtain

$$\overline{\widehat{r}_n} = \frac{r}{n} \widehat{r}_{n \in s^c} + \left(1 - \frac{r}{n}\right) \overline{\widehat{r}_{n-1}} \tag{4.39}$$

$$\widehat{\sigma}_n^2 = \left(\frac{r}{n}\right)^2 \widehat{\sigma}_{n \in s^c}^2 + \left(1 - \frac{r}{n}\right)^2 \widehat{\sigma}_{n-1}^2, \tag{4.40}$$

where $\widehat{r}_{n \in s^c}$ and $\widehat{\sigma}_{n \in s^c}^2$ are defined by (4.37) and (4.38) with $n \in S_m^c$.

By taking $P_y = \sigma_n^2$, and

$$\widehat{P}_y = \frac{r}{dm} \sum_{s \in S_m} \left(H(t_s, \widehat{\theta}_s) - \frac{1}{m} \sum_{s \in S_m} H(t_s, \widehat{\theta}_s) \right) ([\cdots])^T, \tag{4.41}$$

then one can solve for R from (4.18) to get

$$R = \widehat{\sigma}_n^2 - \widehat{P}_y. \tag{4.42}$$

Because the LSQ estimator finds a deterministic realization of each $\widehat{\theta}_s$ assuming no stochastic forcing, it is noted that when using the definitions (4.38) and (4.41), the $\widehat{Q}_y$ term of (4.42) is identically equal to the zero matrix. It is also noted that by combining the results of [87] and [83], both $\widehat{\sigma}_n^2$ and $\widehat{P}_y$ are each aymptotically consistent, and thus R is asymptotically consistent as well.

One can also account for the measurement bias in (4.13) to correct the expected output signal

$$\overline{\widehat{x^+}} = \overline{\widehat{x^-}} + K\left(y - \overline{y} - \overline{\widehat{r_n}}\right). \tag{4.43}$$

4.4.2 Estimating Q from the ensemble filter

When comparing the jackknife LSQ estimator model (4.23) to the SDE model (4.2), the parameter vector θ of the LSQ estimator is usually comprised of the SDE state values $x(t)$ at some time t_k:

$$\theta_k = \begin{pmatrix} x(t_k) \\ \\ x_p \end{pmatrix}, \tag{4.44}$$

where the SDE state values $x(t)$ are augmented by the SDE model parameters x_p having zero deterministic dynamics (i.e., $dx_p = Q_p dw$). It follows from (4.3) that

$$\begin{bmatrix} x(t_{k+1}) \\ \\ x_p \end{bmatrix} = \begin{bmatrix} F(t_{k+1}, \theta_k) \\ \\ 0 \end{bmatrix} + \sqrt{Q_k \Delta t_k} N(0,1). \tag{4.45}$$

Here,

$$Q_k = \begin{bmatrix} Q_t & Q_{tp} \\ \\ Q_{tp} & Q_p \end{bmatrix},$$

where Q_t is the Q from (4.3), Q_p is the auto-covariance of uncertainty in the parameters, and Q_{tp} represents the cross-covariance between uncertainty in the states and parameters. Together, Q_k defines the process uncertainty of the augmented stochastic model (4.45).

By treating each jackknife estimate $\widehat{\theta}_s$ as an ensemble estimate $\widehat{\theta}_i$, we have N ensemble estimates at time t_k:

$$\widehat{x}^{(i)} = F(t_k, \widehat{\theta}_{ki}), \tag{4.46}$$

$$\widehat{P}_x = \frac{1}{N} \sum (\widehat{x}^{(i)} - \overline{\widehat{x}^{(i)}})([\cdots])^T. \tag{4.47}$$

Essentially, the jackknife samples represent an ensemble of state estimates via the transformation of (4.45). In terms of the ensemble filtering framework, the posterior state covariance matrix P_x^+ for state values $x(t_k) = \theta_k$ can be estimated from the jackknife variance estimate v_k, and the prior state covariance matrix P_x^- for the state values $x(t_k) = \theta_{k-1}$ can be estimated by evolving ensemble members backward in time to t_k (similar to the prediction step in UKF) and calculating the ensemble variance at that time step, say $\widehat{P}_x^-$. The justification here is that both P_x^+ and v_n are representations of the state covariance matrix after assimilating new data. By substituting $v_n = \widehat{P}_x^+$ into (4.22) and taking $\widehat{\sigma}_n^2 = \widehat{P}_y + R$, one can explicitly solve for Q:

$$Q = \frac{1}{\Delta t_k} \left(v_n - \widehat{P}_x^- + \widehat{P}_{xy} (\widehat{\sigma}_n^2 - \widehat{Q}_y)^{-1} \widehat{P}_{xy}^T \right). \tag{4.48}$$

4.4.3 Discussion

To simplify the computation of (4.48), the $\widehat{Q}_y$ can be omitted from (4.48), which will yield a pessimistic (i.e., greater in norm) solution for Q since $\widehat{Q}_y$ is positive semi-definite and contributes positively to an inverted term. For many applications, including robust control, this is an acceptable approximation.

For highly nonlinear systems, the LSQ procedure may possibly find a region of minima that are located significantly further away from the dominant mode. These types of secondary modes can quickly emerge and cause the $\widehat{P}_x^-$ to be large enough to make (4.48) negative semi-definite. One solution to this problem would be to implement a Gaussian mixture model (GMM) on the ensemble of realizations and run the adaptive Kalman filter on the constituent normal distributions of ensemble members. In cases where this approach is too computationally costly, the $\widehat{P}_x^-$ term can be omitted from (4.48) to, again, yield an even more pessimistic solution for Q.

4.5 Example application: logistic data

To demonstrate the performance of the adaptive jackknife estimator, we shall consider a simple logistic model with discrete measurements and additive noise:

$$\begin{pmatrix} dx \\ d\beta \\ dN \end{pmatrix} = \begin{pmatrix} \beta x \left(1 - \frac{x}{N}\right) \\ 0 \\ 0 \end{pmatrix} dt + \sqrt{Q} dw \tag{4.49}$$

$$y_k = x(t_k) + \sqrt{R} N(0,1), \tag{4.50}$$

where $\beta \in \mathbf{R}^+$ is the growth parameter, $N \in \mathbf{R}^+$ is the upper bound of x, and dq and $\sqrt{R}N(0,1)$ are the noise processes described in Section 4.2. The logistic model defined by Eqs. (4.49) and (4.50) is a common model used to describe the adoption of a behavior or new technology [17], and is known to have well known convergence properties when

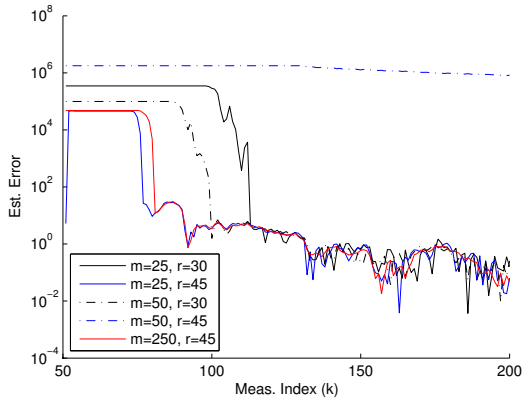


Figure 4.1: Adaptive jackknife estimation performance evaluation for a logistic model, with different jackknife parameter values. In all test cases, $n = 50$ and $\mu = n - r$.

using a jackknife sampling LSQ variance estimator [83]. It is also noted that the integral of the deterministic part of (4.49) (i.e., $dx = \beta x \left(1 - \frac{x}{N}\right) dt$) has the solution:

$$x(t) = \frac{N x(0) \exp \beta t}{N + x(0) \left(\exp \beta t - 1\right)}. \tag{4.51}$$

We simulated a sequence of 200 measurements y_k, at times uniformly distributed on the interval $t = [0, 80]$, with initial values $(x(0), \beta, N) = (1, 0.225, 500)$, and noise covariance matrices $Q = \text{diag}(15, 0.001, 10)$ and $R = 1$. Figure 4.1 shows the error, in Euclidean norm, between the state estimate of the adaptive jackknife filter and the value of (4.51) at time t_k. For a fixed *burn-in* period of 50 measurements, we find that the estimate of the augmented state vector converges with a greater number of included measurements r, and fewer jackknife samples m. With a greater value of r, the adaptive jackknife filter is able to use as many measurements as possible during the burn-in initialization phase, which (i) causes a reduction in the jackknife variance estimate according to (4.32), and (ii) results in an initial estimate closer to the true value by causing the

value r/n to be large. Using fewer jackknife samples (i.e., $m = 25$ vs $m = 50$) seems to also counter-intuitively produce a fast convergence result in this example, because few jackknife samples are needed to accurately represent the uncertainty distributions in the model. Choosing $m = 50$ causes an over-sampling of outliers, and it is not until we have $m = 250$ that the true distribution emerges.

4.6 Conclusion and Future Work

We have shown how one can implement the techniques of jackknife variance estimators as they apply to least squares estimators to construct an adaptive, nonparametric, and computationally efficient statistical nonlinear filter.

One issue that we left as an assumption is that for each jackknife estimate, there exists a solution to the LSQ problem. In fact, this is not a far-fetched assumption to make because bootstrap methods (similar to jackknife sampling) have been shown to efficiently search for the solution to the general LSQ problem [89]. Lastly, we also remark that jackknife sampling LSQ problem is easily broken down to a parallel computation problem, since the LSQ solution for each jackknife sample of the data can be solved independently of each other jackknife sample. Therefore, there is room for future work on this subject to increase computational efficiency, both with respect to improving LSQ estimation and parallelizing each step of the adaptive algorithm.

At first glance it seems as though one can use the sum of least-squares of residuals (i.e., the least-squares objective function) as an estimate for the signal variance. However,

this variance estimate is one point sample of the variance estimate for the available data, and does not say anything about the confidence level of this estimate with respect to the true variance.

Chapter 5

Does the network model explain the measured data?

With a known network topology, the agent-based adoption model of Chapter 2 describes the simple contagion behavior of information on that social network, where each individual on the network can only become informed from another member of the same network. Furthermore, the results of Chapter 2 are careful to only discuss the agent-based model as a tool for investigating the effects of network structure on the rate of information propagation when controlling for all other parameters. Suppose that we know the other parameters in addition to knowing the network structure, and let us suppose also that we have actual time series data for the adoption of a behavior on a social network. *How well do these agent-based network models explain real data?*

The statistical tools of Chapter 4 allow us to estimate the underlying stochastic model that produced an observed realization of the data. To do this, we estimated a mean-field solution (LSQ jackknife mean) and variance about that mean-field solution (LSQ jackknife variance) of the adoption data. The ensemble of mean-field solutions from the

jackknife sampling procedure can be described by their distribution of parameters and initial conditions, and these same ensemble parameters and initial conditions can also be used as the parameters and initial conditions for the agent-based network model. Since the agent-based network model has the same parameter and initial value distribution as the stochastic data model, a paired statistical hypothesis test can be used to determine how likely the agent-based network ensemble realizations were generated by the stochastic data model, when a specified null hypothesis is true. Ultimately, our findings will be used to motivate the need for a better way of modeling agent behavior on social networks, which is discussed in Chapter 6.

5.1 Generating Ensemble Realizations from Data

With repeated experiments, uncertainty in the data can be estimated by an ensemble distribution of the data. However, with real data from the Twitter social network, adoption phenomena can only be observed once. For example, once a user adopts a hashtag, it is very unlikely that they will forget the hashtag and re-adopt it in the same context. Not only does it remain an open research problem for how to reliably detect when someone on a social network has forgotten a parcel of information, it still remains a challenge to control for all other variables that affect the state of the social network. Fortunately, all is not lost. The adaptive jackknife sampling tools discussed in Chapter 4 provide a way for us to estimate the stochastic model that generated each individual

hashtag adoption time series. The solution to the SDE of the stochastic model can then be solved via Monte-Carlo simulation [90].

Since the methods of Chapter 4 tend to over estimate the magnitude of the stochastic forcing term $\sqrt{Q}$ of (4.1a), a more reliable estimate of $\sqrt{Q}$ can be found by *simulated maximum likelihood* (SML) [91]. The open source SDE Toolbox for Matlab [92] implements the SML algorithm outlined in [91], and uses the Euler-Murayama method [90] to Itō integrate each ensemble realization in the Monte Carlo simulated SDE solution. The SDE Toolbox also implements Matlab's built-in Nelder-Mead simplex (direct search) method [93], which is used to find the $\sqrt{Q}$ value that maximizes the maximum likelihood objective function of the SML algorithm. Therefore, the adaptive jackknife procedure finds the maximum likelihood parameters (in a LSQ sense) that define the drift term of (5.1), and the SML algorithm finds the magnitude of the stochastic forcing (diffusion) term.

After the jackknife procedure finds the maximum likelihood estimates of the $I(0)$, N, and I_{t-} parameters, we treat these values as constants. We also assume, in this chapter, that the data measurements are direct measurements of the I state variable (i.e., $R = 0$), and obtain the 1-D SDE:

$$dI = \beta I \left(1 - \frac{I}{N}\right) dt + \sqrt{Q} dw \tag{5.1a}$$

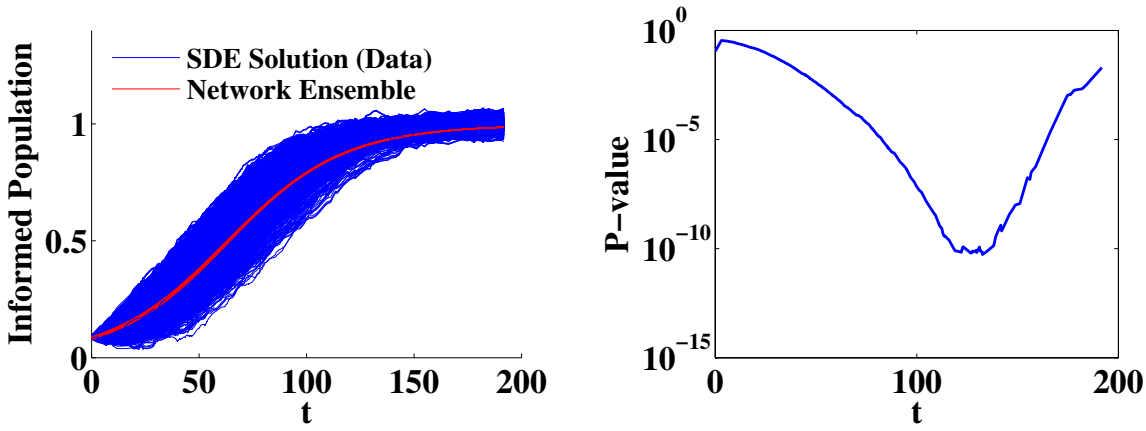
$$y = I + I_{t-}. \tag{5.1b}$$

The SML algorithm is used to find the maximum likelihood estimate of $\sqrt{Q}$ in (5.1).

Using the notation established in Section 4.3, the adaptive jackknife scheme is implemented on each hashtag's observed time series with $m = 250$ and $r = 4n/5$, which are sufficient values for estimating the logistic equation according to the results of Section 4.5. Because a sufficient number of data points is required to do the adaptive jackknife estimation, we include only those hashtags that have at least $n = 30$ data points. Therefore, for each of these hashtags, we chose $\mu = 4$ and used the first $n_0 = r + 5$ data points to begin the adaptive estimation procedure. The adaptive jackknife estimation was also implemented on the ICB computer cluster at UCSB, where we had reserved access to exactly 25 nodes via the built-in Matlab parallel computation toolbox.

After finding the jackknife parameter estimates for the hashtags with a sufficient amount of data, complexity of the SML algorithm limited it to only being able to converge to a solution for a subset of the remaining hashtags. In total, we were able to accurately estimate all unknown parameters for 28.6% of the hashtags in our data set. Although it remains as future work to accurately and robustly estimate all parameters for a given SDE from real data, generalizing a computationally tenable procedure for this task is an ongoing research problem that is outside the scope of this manuscript.

We remark that having access to only 25 compute nodes limited us to $m = 25$ realizations of the agent-based model because a single realization on one processor took on the order of 6 minutes to 18 hours, depending on the number of data measurements. Although we could have increased the value of m at the cost of linearly increasing total compute time, the findings of Section 4 demonstrate that $m = 25$ is sufficient. It is also



(a) (b)

Figure 5.1: (a) Ensemble realizations normalized by population size for the respective model, and (b) time-dependent p-values for the *#nobama* hashtag.

noted that each parameter set of these realizations was chosen from the set of parameters corresponding with one of the 250 LSQ jackknife ensemble members. For each of these 25 randomly selected parameter sets, a set of 1000 Monte Carlo ensemble realizations was simulated to produce the 25000 realizations that compose the empirical data density distribution (see Figure 5.1a for example).

5.2 Model Comparison

Because the carrying capacity (i.e., relevant population size) of the agent-based network model (2.15) may be different from the carrying capacity of the SDE (5.1a) that best explains the data, we normalize each model by its respective carrying capacity for each hashtag. We shall denote $I_N(t)$ to represent the size of the informed population of the agent-based network model at time t, and, similarly, $I_D(t)$ to represent the size of

the informed population represented by the data and included in the SDE data model. As mentioned in [17], the carrying capacity of the logistic model is an asymptotically stable fixed point. For the network model, the carrying capacity of the hashtag's topical backbone network, say $I_N(\infty)$, is determined by the set of users who are reachable from the set of known infected individuals (as given by the data realization for that hashtag), while the carrying capacity for the data, denoted as $I_D(\infty)$, is a parameter found by the adaptive jackknife procedure of Chapter 4 (i.e., $I_D(\infty) = N$). Therefore, each user in the agent-based simulation was initialized with the same likelihood of being informed so that both the agent-based network simulation and the data Monte Carlo realizations have the same initial conditions in their respective normalized scales (i.e., $I_N(0)/I_N(\infty) = I_D(0)/I_D(\infty)$).

By construction of this experiment, the normalized homogeneous approximation of the difference equation (2.15) is identical in the continuous limit ($h \to 0$) to the drift term of (5.1a), when normalized with respect to $I_D(\infty)$. Both systems have the same β value, and the same initial likelihood of informed individuals. The normalized agent-based network model and the normalized SDE data model differ by their additive terms. The additive term of the agent-based network model is the bound on the heterogeneous network effects, while the additive term of the SDE data model is the stochastic forcing (diffusion) term. In this section we investigate the significance of these additive terms by determining how well the uncertainty in the data model (stochastic forcing term) is explained by the heterogeneous effects of the network (homogeneous approximation

error). Section 5.2.1 uses a statistical hypothesis test to investigate the probability that the agent-based network realizations were generated by the same density function that generated the data. Section 5.2.2 explains the results of Section 5.2.1 in terms of the magnitudes of the additive terms with respect to each other.

5.2.1 Statistical hypothesis test

When a sample statistic is drawn, such as a sample mean, one would like to be able to determine how likely it is that the sample statistic was drawn from a given distribution rather than by random chance. To do this we can perform a statistical hypothesis test on the sample statistic.

For a given point in time t, let us take the sample mean $\mu(t)$ of sample size $m = 25$ as our test statistic. At each time step, we shall employ a *z-test* [94] to determine the likelihood that the sample mean of the agent-based network realizations came from the sample distribution of Monte Carlo ensemble members that comprise the data SDE solution.

The jackknife members define the agent-based ensemble distribution such that the ensemble mean of the network realizations is

$$\mu_N(t) = \frac{1}{m} \sum_{\theta_i} I_N(t), \tag{5.2}$$

for each LSQ jackknife parameter estimate θ_i at time t. Likewise, a sample mean of m randomly chosen ensemble members $I_D(t)$ from the data SDE solution is defined as

$$\mu_D(t) = \frac{1}{m} \sum I_D(t). \tag{5.3}$$

When conducting the z-test, we take the null hypothesis to be

$$H_0 : \mu_N(t) = \mu_D(t). \tag{5.4}$$

It is noted that the distribution of $\mu_D(t)$ can be approximated from the mean and variance of the empirical distribution of $I_D(t)$ values at time t, as implemented by Matlab's built-in `ztest()` function, which was used to produce the results in this section.

Each z-test, conducted at each time step, will produce a *p-value* [94]. The p-value is commonly interpreted as representing the probability that an observation of the sample statistic occurred by random chance under the null hypothesis. However, we shall only consider the p-value for the much simpler task of detecting a binary outcome. For instance, it is often accepted in literature that a p-value of less than 0.95 is enough evidence to reject the null hypothesis [94].

By construction, it must necessarily be true that $\mu_D(0) = \mu_N(0)$ in both the absolute and normalized coordinates. Figure 5.1a shows a set of ensemble realizations in the normalized coordinates, and Figure 5.1b shows how the p-values change in time. Since the normalized initial conditions of the network and data models are the same value, and both models asymptotically approach the unity value in their respective population-normalized coordinates, one should expect the greatest separation between the two models to be

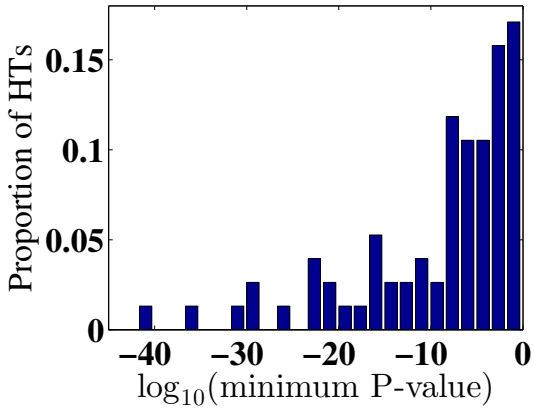


Figure 5.2: Minimum P-values over time for all available hashtags.

greatest during the transient phase of the system. From the example shown in Figure 5.1a, we find that the p-values are closest to $p = 1.0$ where the functions intersect, and near zero at time $t = 120(days)$. Since the minimum p-value for the *#nobama* hashtag is significantly less than 0.95, this indicates that the two ensembles are highly unlikely to be statistically different by random chance.

We repeated this statistical hypothesis test at each time step for each hashtag in our dataset, and recorded the minimum observed p-value in each hashtag's time series. Twenty tree hashtags were discarded because their $\sqrt{Q}$ values were large enough to stochastically drive the data ensemble mean to the $I_D(t) = 0$ (unstable) fixed point, which violates the assumption that no members of the population are able to forget about the hashtag. Figure 5.2 shows the distribution of minimum p-values over the set of remaining ninety eight available hashtags from the original data set. The maximum p-value recorded in Figure 5.2 is 0.68, and suggests that none of the agent-based network ensembles were generated by the same process that generated the data.

5.2.2 Stochastic forcing or network effects?

Let us consider the respective network and data models. The proportion of informed agents in a network defined by the row stochastic adjacency matrix A, is generally described by dividing (2.15) by the network carrying capacity $I_N(\infty)$:

$$\frac{I_{t+h}}{I_N(\infty)} = \frac{I_t}{I_N(\infty)} + h\beta \frac{I_t}{I_N(\infty)} \left(1 - \frac{I}{I_\infty}\right) + O\left(\frac{h \left\|A - R_1\right\|_2}{I_N(\infty)}\right), \tag{5.5}$$

which has an upper bound on the homogeneous approximation error term, as described in Appendix B.3:

$$O\left(\frac{h\|A - R_1\|_2}{I_N(\infty)}\right) \leq O\left(\frac{h\beta_N \|A\|_2}{I_N(\infty)}\right). \tag{5.6}$$

Thus, combining expressions (5.5) and (5.6) gives

$$\frac{I_{t+h}}{I_N(\infty)} = \frac{I_t}{I_N(\infty)} + h\beta \frac{I_t}{I_N(\infty)} \left(1 - \frac{I}{I_\infty}\right) + O\left(\frac{h\beta \|A\|_2}{I_N(\infty)}\right). \tag{5.7}$$

To further simplify our analysis in this section, we can take advantage of the fact that $\beta << 1$ is true for all hashtags in our data set so that (5.7) is numerically stable [95]. Therefore, we can let $x_N(t) = I_t/I_N(\infty)$ and use the continuous ODE:

$$\frac{dx_N}{dt} = \beta x_N \left(1 - x_N\right) + O\left(\frac{h\beta \|A\|_2}{I_N(\infty)}\right) \tag{5.8}$$

as a close approximation of (5.7), since (5.7) is a forward Euler approximation of (5.8). Similarly, we can normalize (5.1a) by $I_D(\infty)$ and let $x_D(t) = I_D(t)/I_d(\infty)$ to obtain:

$$\frac{dx_D}{dt} = \beta x_D \left(1 - x_D\right) + \frac{\sqrt{Q}}{I_D(\infty)} dw. \tag{5.9}$$

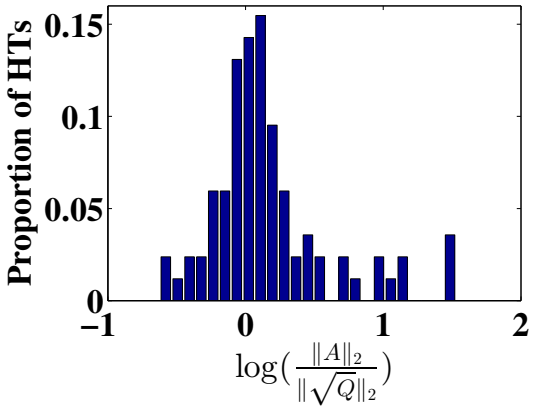


Figure 5.3: Distribution of $\|A\|_2/\|\sqrt{Q}\|_2$ values for the hashtag dataset.

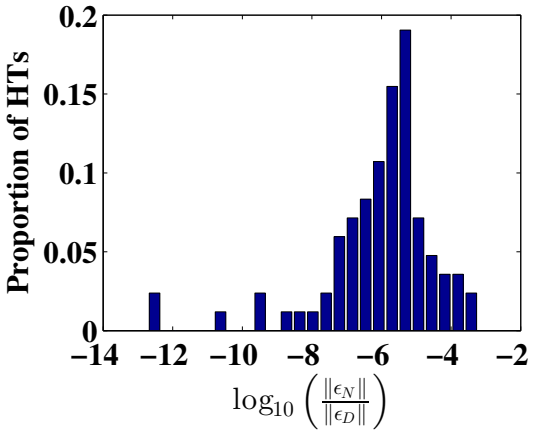


Figure 5.4: Distribution of the normalized ϵ_D vs. normalized ϵ_N values for the hashtag dataset.

When comparing (5.8) to (5.1), the stochastic forcing term

$\epsilon_D \triangleq (\sqrt{Q}/I_D(\infty))dw$ of (5.9) and the $\epsilon_N \triangleq (\beta\|A\|_2/I_N(\infty))$ term of (5.8) are analogous to each other, and we can compare their relative magnitudes to each other. If the magnitudes are relatively close, then it may be possible that the data realizations are actually consistent with each other even though their estimated ensemble behavior is significantly different. However, the evidence shown in Figure 5.4 weakly supports the

null hypothesis (5.4), and leads one to conclude that uncertainty in the data is not explained by the homogeneous approximation error of the network model.

Figure 5.3 shows the distribution of the $\|A\|_2/\|\sqrt{Q}\|_2$ values for the available hashtags in our dataset, and shows that these quantities are of the same order of magnitude. However, since the relative magnitudes of the normalized ϵ_D and ϵ_N values, as shown in Figure 5.4, are often orders of magnitude different, it seems very unlikely that the stochastic forcing term is explained by the homogeneous approximation error for an arbitrary hashtag. These findings support the earlier claims of Section 5.2.1.

5.3 Discussion

Based on the evidence presented in Section 5.2, it is unlikely that the fluctuations in the data are explained by the given network structure. This suggests that if one wished to use an agent-based network model to describe hashtag adoption data within a population, then a more detailed and time-varying edge weighting scheme may be required to sufficiently model each user-user interaction in the agent-based network. With a social network as large as Twitter, which has over 1.47×10^9 edges [50], the procedure one chooses to model the behavior at each edge can quickly become untenable. If this type of procedure were actually tenable, the uncertainty of the edge weight estimations would collectively result in a high level of uncertainty in the overall estimate of the size of the informed population. In contrast, when considering the population model directly,

one may think of the fluctuations in the edge weight functions as canceling each other out so that there is more certainty in the population estimates inferred from the data.

Furthermore, as future work, it may be possible to explain the hashtag adoption data using an SEI-type contagion model [17], which models an exposed (E) class of individuals who are neighbors of hashtag users but have not used the hashtag themselves. One can compare the estimated size of E-class individuals to the number of exposed neighbors in the network model using our adaptive jackknife estimation strategy, and compare the size of these sets in order to gain more insight into the network effects of hashtag adoption. However, population models that are more complex than the simple SI systems become even more complex when constructing their agent-based network analogue since complexity of the agent-based model scales with the size of the network. Once the agent-based model is defined for these more complex population models, it still remains a challenge how one can reduce the agent-based network equations to a coarse population model for comparison.

It is also known that the value of social information in a network varies in time. For instance, it has been empirically shown that the recency of a news story affects its rate of imitation among news sources [39]. The authors of [39] suggest a parameter, say η, that describes the global proportion of news sources writing about a given meme at time t after its first mention:

$$\frac{d\eta}{dt} = cq\eta t^{-1},$$

where c is a normalization constant, and q is a constant rate parameter. If we were to perceive the aggregate discussion of hashtag related stories in news outlets as a proxy for the general interest in a given meme, then the value of η represents a scaling of the population's hashtag adoption curve $I(t)$:

$$\begin{pmatrix} \frac{dI}{dt} \\ \\ \frac{d\eta}{dt} \end{pmatrix} = \begin{pmatrix} h\eta\beta I\left(1 - \frac{I}{I_\infty}\right) \\ \\ cq\eta t^{-1} \end{pmatrix}. \tag{5.10}$$

Since η scales the adoption rate β in (5.10), the ability of the agent-based network model to track the data will not be affected, but only relatively scaled. One could add more parameters to the existing model to account for various population effects, however, it is unlikely that these types of modifications will lessen the disparity between the population model and the agent-based network model since population effects will affect the mean behavior of both models equally.

At this point, it is clear that population-level models are better for predicting population-level behavior than the agent-based network models. At the individual level, is it possible to make better predictions about who will adopt a hashtag conditioned on their neighbors who have already used the hashtag? The evidence suggests that the agent-based network model of (5.5) seems to be ill suited for this task. How can we do better?

Chapter 6

A Better User Model

Instead of using the agent-based network model for predicting global hashtag adoption, which is outperformed by the optimal estimation techniques outlined in Chapter 4, we propose a novel user model that is capable of capturing topic specific behavior and is capable of generally predicting adoption behavior at the agent level. Since the jackknife methods already optimally estimate global hashtag behavior from the data, we shall investigate the efficacy of a genetically inspired user model that is able to describe user behavior with respect to hashtags of different topics.

Trends and influence in social media are mediated by the individual behavior of users and organizations embedded in a follower/subscription network. The social media network structure differs from a friendship network in that users are allowed to *follow* any other user and follower links are not necessarily bi-directional. While a link enables a possible influence channel, it is not always an active entity, since a follower is not necessarily interested in all of the content that a *followee* posts. Furthermore, two individuals are likely to regard the same token of information differently. Understanding how infor-

mation spreads and which links are active requires characterizing the users' individual behavior, and thus going beyond the static network structure. A natural question then arises: *Are social media users consistent in their interest and susceptibility to certain topics?*

In this chapter, we answer the above question by demonstrating a persistent topic-specific behavior in real-world social media. We propose a user model, termed *genotype*, that summarizes a user's topic-specific footprint in the information dissemination process, based on empirical data. The social media genotype, similar to a biological genotype, captures unique user traits and variations in different genes (topics). Within the genotype model, a node becomes an individual represented by a set of unique invariant properties.

For our particular analysis, the genotypes summarize the propensity and activity level in adoption, transformation, and propagation of information within the context of different topics. We propose a specific set of properties describing the adoption and use of topic-specific Twitter hashtags. The model, however, applies to more general settings capturing, for example, dissemination of URLs or sentiments in the network.

We construct the genome (collection of user genotypes) of a large social media dataset from Twitter, comprised of both follower structure and associated posts. The existence of stable genotypes (behavior) leads to natural further questions: *(1) Can this consistent user behavior be employed to categorize novel information based on its spread pattern? (2) Can one utilize the genotypes and the topic-specific influence backbone to (i) predict likely adopters/influencers for new information from a known topic and (ii) improve*

the network utility by reducing latency of disseminated information? In this chapter, we explore the potential of the genotype model to answer the first question within the context of Twitter[1].

To validate the consistency of genotypes, we show that combining genotype-based classifiers into a composite (network-wide) classifier achieves accuracy of 87% in predicting the topic of novel hashtags that spread in the network. We extract and analyze topic-specific influence backbone networks and show that they structurally differ from the static follower network. When considering the population level dynamics, using a simple contagion model, we show that hashtags of a known topic propagate at the greatest rate on backbone networks of the same topic, and that this result is consistent with the local user model.

We remark that the data values for each genotype metric are likely to be affected by the fact that 80% of the SNAP users' messages were not recorded. In addition, not all hashtags we encounter can be attributed to a topic. Nonetheless, all metrics in this study are affected equally, and evaluated relative to each other. Obtaining complete snapshots of network structure at any given point in time in these experiments is untenable. Thus, we acknowledge this limitation and cast our results in the context of only what is known about the network structures and posts within the Twitter dataset.

[1]Both questions are discussed in [4], and the work pertaining to question (2) is primarily attributed to Petko Bogdanov.

6.1 Related Work

The network structure has been central in studying influence and information dissemination in traditional social network research [96, 97]. Large social media systems, different from traditional social networks, tend to exhibit relatively denser follower structure, non-homogeneous participation of nodes, and topic specialization/interest of individual users. Twitter, for example, is known to be structurally different from human social networks [50], and the intrinsic topics of circulated hashtags are central to their adoption [52].

A diverse body of research has been dedicated to understanding influence and information spread on networks, from theories in sociology [98] to epidemiology [17, 33], leading to empirical *large-scale* studies enabled by social web systems [49, 52, 99, 100]. Here, we postulate that the influence structure varies across topics [57] and is further personalized for individual node pairs. Lin and colleagues [101] also focus on topic-specific diffusion by co-learning latent topics and their evolution in online communities. The diffusion that the authors of [101] predict is implicit, meaning that nodes are part of the diffusion if they use language corresponding to the latent topics. In contrast, we focus on topic-specific user genotypes and influence structures concerned with passing of observable information tokens and their temporal adoption properties.

Earlier data-centered studies have shown that sentiment [99] and local network structure [52] have an effect on the spread of ideas. The novelty of our approach is the focus on content features to which users react. Previous content-based analyses of Tweets have

adopted latent topic models [102, 103]. We tie both content and behavioral features to the network's individuals.

With regard to influence network structure and authoritative sources discovery, Rodriguez and colleagues [104] were able to infer the structure and dynamics of information (influence) pathways, based on the spread of memes or keywords. Bakshy et al. [105] focus on Twitter influencers who are roots of large cascades and have many followers, while Pal et al [106] adopt clustering and ranking based on structural and content characteristics to discover authoritative users. Although the above works are similar to ours in that they focus on influence structures and user summaries, our genotype targets capturing the invariant user behavior and information spread within topics as a whole, involving a collection of topically related information parcels.

Our framework is inspired by biology and evolution, similar to Reali and Griffiths [107]. We broaden the genotype interpretation beyond word variants, and demonstrate their predictive utility. Our goal is to treat the observable content as a genetic parcel of information that users pass on to one another, while potentially introducing a delay or alteration to the message. An added benefit of this approach is that similarity of behavior toward certain types of messages among users may indicate social affinity (of interests, attitudes, etc.), provide important information about transmission paths in the network, and predict future edge formation [108].

6.2 Genotype Model

Here we define our genotype model capturing the topic-specific behavior of a single user (node) within a social media network. Our main premise is that, based on observed network behavior, we can derive a consistent signature of a user. Hence, the genotype model is an individual user model, by definition, in the sense that it represents the behavioral traits of a social network user. For our analysis, the genotype captures adoption and reposting of new information, activity levels, and latency of reaction to new information sent by influential neighbors. Other behavioral traits can be incorporated as well. The genotype is topic-specific as we summarize the behavioral traits with respect to a set of predefined topics.

Recall that a social media network $N(U, E)$ is a set of users (nodes) U and a set of follow links E. A directed follow link $e = (u, v), e \in E$ connects a source user u (*followee*) to a destination user v (*follower*). The network structure determines how users get exposed to information posted by their followees. The static network does not necessarily capture influence as users do not react to all information to which they are exposed. To account for the latter, we model the behavior of individual users taking into account their context in the follower network.

In its most general form, a user's genotype G_u is an entity embedded in a multi-dimensional feature space that summarizes the *observable behavior* of user u with respect to different topics. It is up to the practitioner to define the different dimensions of the topic feature space and the relevant aspects of observable behavior in the network locality

of a node. Each genotype value can be viewed as an allele that the user introduces to the process of message propagation through a network.

In our study, we focus on hashtag usage within Twitter, since hashtags are simple user-generated tokens that annotate tweets generated by either a social group or designating a specific social phenomenon, and are often "learned" from others on the social network [51]. In this context, a hashtag serves as a genetic parcel of cultural information, just like alleles of a gene within a biological context. Hashtags can be associated with topics such that an individual's response to a collection of hashtags within a topic indicates a user's propensity to respond to other hashtags within that same topic.

We consider a finite set of hashtags $H = \{h\}$, each associated with a topic $T_i \in T$. To obtain the genotype, we analyze the social media message (tweet) stream produced by a user u, with respect to H. Let us define $m(\cdot)$ to be a function that maps each occurrence of (u, h) to a real values $m : \{(u, h)\} \mapsto \mathbf{R}$. The set of hashtags associated with topic T_i and adopted by user u are denoted as $H_{(u,T_i)} := \{h\}_{T_i} \cap \{h\}_u$, where $\{h\}_{T_i}$ is the set of hashtags in topic T_i and $\{h\}_u$ is the set of hashtags adopted by user u. The i^{th} element of the user genotype G_u is the set of $\{m(u, h) \mid h \in H_{(u,T_i)}\}$ values. We remark that this set of values may also be reduced to their average value or some approximated distribution function if one wishes to have a coarser representation of the data.

To construct each user's topic-genotype from empirical data, we consider a variety of metrics $m(\cdot)$ for (u, h) pairs, listed in Table 6.1. These metrics serve the purpose of quantifying a user's response to a hashtag by defining the data values that are used to

Metric	Function definition	Notes
Time	TIME$(u,h) = min_{(u,h)}(t(u,h)) - min_{v \in V_u}(t(v,h))$, where $t(u,h)$ is the time (u,h) occurs and V_u is the set of followees of u.	The absolute amount of time between a users first exposure to the given hashtag and his first use of that same hashtag.
Number of Uses	N-USES$(u,h) = \lvert\{(u,h)\}\rvert$, where $\lvert\cdot\rvert$ is the cardinality function.	The total number of occurrences of the (u,h) pair.
Number of Parents	N-PAR$(u,h) = \lvert\{v \in V_u \mid t(v,h) < t(u,h)\}\rvert$	The number of followees to adopt before the given user.
Fraction of Parents	F-PAR = N-PAR$(u,h)/\lvert V_u\rvert$.	The fraction of a user's followees who have adopted the hashtag prior to the user.
Latency	LAT$(u,h) = (\lvert\{h_j \in H_{T_i} \mid H_{T_i} \ni h$, and $t(u,h_j) < t(u,h)\}\rvert)^{-1}$.	The inverse of the number of same-topic hashtags posted to the user's time-line between his first exposure to the hashtag and his first use of the hashtag.
Log-latency	LOG-LAT$(u,h) = log\,(\text{LAT}(u,h)/Avg(\text{LAT}(w,h)$ s.t. $w \in U))$.	The logarithm of each latency value after each latency value has been divided by the mean latency value for that hashtag.

Table 6.1: Behavior-based metrics that are components of the topic-specific user genotype.

estimate the topic distributions. While TIME and N-USES are intuitively obvious metric choices, LAT and LOG-LAT are novel to this manuscript. N-PAR and F-PAR have been previously studied in a different context [52], and are included here for comparison.

While we define the user genotypes based on adoption of hashtags in Twitter similar models can be built in other networks as well. The follower network structure in Twitter forms a directed graph and hence the definition can be easily generalized to undirected networks such as those of systems like Facebook and Google+. Instead of hashtags one

can focus on other aspects of behavior such as adoption of new phrases, hyper-links or other tokens that carry topical information.

6.3 Genotype model validation in Twitter

To justify the genotype model as a meaningful representation of social network users, we demonstrate that it is capable of capturing stable individual user behavior for a given topic. We seek to evaluate the stability of configuration of multiple users' genotype values within a topic, and use a classification task and the obtained (training/testing) accuracy as a measure of consistency for our genotype model. Within this context, we compare different genotype dimensions and evaluate the level to which each of them captures characteristic invariant properties of a social media user.

6.3.1 Topic consistency for individual users

Our hypothesis is that individual users exhibit consistent behavior of adopting and using hashtags (stable genotype) within a known topic. If we are able to capture such invariant user characteristics in our genotype metrics, then we can turn to employing the genotypes for applications. We compute genotype values according to our collection of hashtags with known topics by training a per-user Linear Discriminant (LD) topic classifier to learn the separation among topics. The LD algorithm fits a multivariate normal density (via the standard EM algorithm) to each group with a covariance estimate that is assumed to be equal for each topic [109], and was implemented via Matlab's

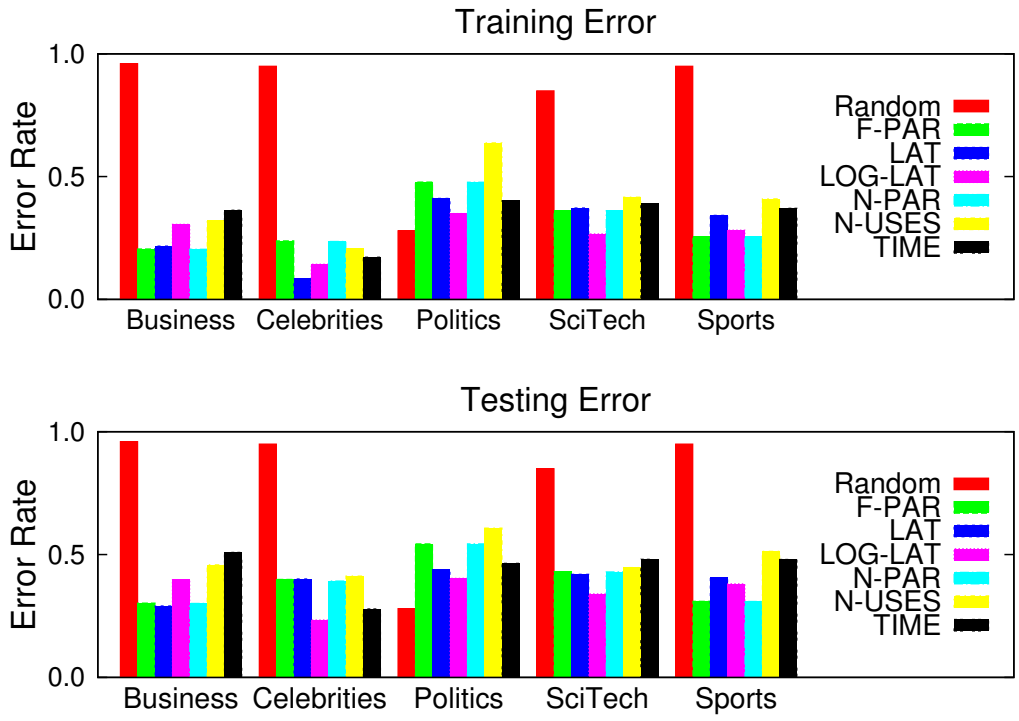


Figure 6.1: Training and testing accuracy of hashtag classification in a leave-one-out Linear Discriminant classification.

`classify()` function. Since each metric in our study is a scalar real value, the LD classifier for each user partitions the real line into adjacent convex sections corresponding with each topic of maximum likelihood. Therefore, assigning a hashag to a topic for a specific user becomes a simple binary decision, where we place the hashtag on the real line according to it's metric value and note whether or not it is assigned the correct topic label.

For our application, consider the LOG-LAT genotype metric: for a user u, we have a set of observed LOG-LAT values (based on multiple hashtags) that are associated with the corresponding topics. If the user u is consistent in her reaction to each topic, then the

LOG-LAT values per topic will allow the construction of a classifier with low training and testing error. It is also noted that each hashtag does end up having a topic distribution, but for the scope of this study, a sufficient hashtag classification should at least agree in the topic of greatest probability/likelihood, which is what is presented here. Moreover, to be able to estimate probability distributions for each topic, we only consider users who have at least two hashtag uses in each topic.

The consistency of user responses is evaluated using a leave-one-hashtag-out validation. Given the full set of (u, h) response values, we withhold all pairs including a validation hashtag h and employ the rest of the pairs involving hashtags of known topic to estimate the individual user's topic genotype. We repeat this for all genotype metrics. The training and testing error rate for this experiment are presented in Figure 6.1, and their similar error rates demonstrate how consistent users are at classifying hashtags into topics. In both cases, our genotype metrics enable significantly lower error rates than a Random model (i.e. random prediction based on number of hashtags within a topic), demonstrating that, in general, genotype metrics capture consistent topic-wise behavior. One exception is the Politics topic as it has comparatively many more hashtags than other topics, skewing the random topic distribution resulting in slightly lower error. Across genotype metrics, we observe that normalized latency of adoption (LOG-LAT) is more consistent per user than alternatives.

	Bus.	Celeb.	Poli.	Sci./Tech.	Sport	E[x]
Random Error	0.96	0.95	0.28	0.85	0.95	0.45
F-PAR	0.50	0.88	0.61	0.15	0.09	0.41
LAT	0.09	0.46	0.18	0.19	0.25	0.21
LOG-LAT	0.05	0.13	0.19	0.12	0.03	0.13
N-PAR	0.09	0.50	0.88	0.09	0.03	0.40
N-USES	0.45	0.42	0.90	0.22	0.56	0.54
TIME	1.0	1.0	0.01	0.92	0.88	0.61

Table 6.2: Error rates of the NB consensus topic classification. $E[x]$ is the expected error across topics.

6.3.2 Topic consistency within the network

In order to track topics, or recommend relevant content, it is essential to understand the topic of newly-arising hashtags. To this end, we leverage the existing genotypes for the SNAP dataset and build a concensus classification framework based on how new hashtags spread within the network of genotype-annotated nodes (i.e., Twitter users). We begin by using the individual user classifications from the validation set of hashtags, and then implement a Naive Bayes (NB) algorithm to achieve consensus on the topic classification of each validation hashtag. Additionally, we also demonstrate that consensus becomes more accurate as more individuals use the given hashtag. While individual users may exhibit some inconsistencies in how they behave with respect to hashtags within a topic, an ensemble of users' genotypes remains more consistent overall. To demonstrate this effect, we extend our classification-based evaluation to the network level. We implement a network-wide ensemble-based Naive Bayes (NB) classifier that combines output of individual user classifiers to achieve network-wide consensus on the topic classification of each validation hashtag.

To implement a NB consensus classifier on the output of each user's local LD classifier, posterior topic distributions from the LD classifier are required for each topic of each user's genotype. Since the LD classifer optimally fits a multinomial normal distribution to the sets of topics for each user, one can use this multinomial distribution to estimate the posterior likelihood that a newly classified hashtag (i.e., from the validation set) belongs to a specified topic. However, the LD classifier assumes the same covariance estimate for each topic, which causes the posterior likelihood to be underestimated for tightly clustered hashtags of the same topic, and over estimated for relatively dispersed clusters of same-topic hashtags. To correct for the uniform covariance assumption, posterior likelihood estimates are calculated from the empirical hashtag distributions of each topic (for the specified user). To remain consistent with the normality assumption of the LD classifier, we assume normality for these empirical distributions within each topic, where the mean values are centered about the correctly classified training hashtags and the variance is computed from all training hashtags for that topic.

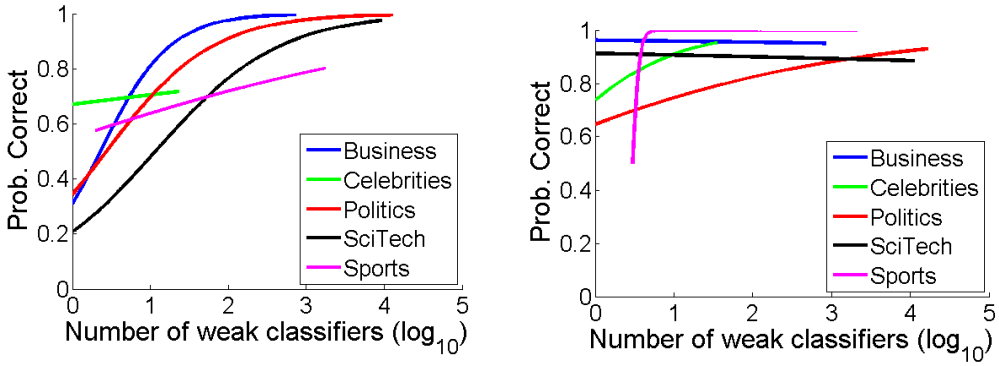
With regards to our NB implementation, the topic prior distributions are estimated from the relative proportion of hashtags in each topic, and the hashtag's ultimate topic classification is determined by the maximum posterior likelihood over the network (all user-wise LD classification outputs).

We note that since each hashtag was used by a moderate subset of users (compared to the size of the whole network), only those users' genotypes were needed for training the classifiers. This locality of hashtag usage, and hence relevant users, is computationally

advantageous because the amount of data needed for classification is bounded by the number of users who used the validation hashtag, and the complexity of their genotypes. However, the inherent data sparsity may become a disadvantage by limiting the classification to binary (in-topic or not-in-topic as opposed to multi-class) and degrading the overall network classification performance when too few local classifiers are available.

Table 6.2 summarizes the testing error rate of our NB scheme for classifying hashtags into topics in a leave-one-hashtag-out validation. The consensus error rate decreases compared to local classifiers (Figure 6.1), demonstrating that the genotypes, as a complex, are more stable and consistent than individual users. The lowest error rate of 0.13 is achieved when using the LOG-LAT metric. The TIME metric happened to be the least accurate metric of them all, because individual user response time values (TIME) showed the least discernable clustering behavior. The accuracy of the TIME metric performed most similar to the null (Random) model when compared the other metrics on a topic-by-topic basis, but TIME happened to be more biased towards political hashtags because they occurred most frequently in the dataset.

The latency genotype metrics that are most invariant (LAT and LOG-LAT) implicitly normalize their time scales of response with respect to the user's own frequency of activity, which is a feature not captured by the absolute TIME metric, or any of the other metrics. Furthermore, both of these metrics incorporate the network structure, measuring the message offset since the earliest exposure to the hashtag via a followee. LOG-LAT has a slight advantage over LAT because it suppresses the background noise



(a) LAT Net Classifier (b) LOG-LAT Net Classifier

Figure 6.2: Accuracy of the network classification as a function of the number of local classifiers (SNAP). A logistic function is fit to each topic's accuracy.

of each hashtag measurement. However, LOG-LAT has the disadvantage of being dependent on a network-wide latency measurement for the same hashtag, which might be harder to obtain in practice. In this sense, LAT is a more practical genotype dimension when summarizing individual user behavior in real time.

While the system of all user genotypes exhibits significant consistency (high classification accuracy), it is useful to know how many user genotypes are needed to obtain a good classification (i.e., detect a network-wide topic-specific spread). We observe an increasing classification accuracy with the number of users included in the NB scheme. Figures 6.2a and 6.2b show the dependence of accuracy on number of local LD classifiers included per topic. All curves increase sharply, indicating that variability within individuals is easily overcome by considering a small subset of users within the network. In fact, the Business and Sci./Tech. accuracies in Figure 6.2 are most accurate for the smallest subset of users (i.e., fewest number of local classifiers), and then decrease slightly as less reliable

individuals are included in the network classifier. Overall, the accuracy of the LOG-LAT network classifier tends to increase *faster* to its optimal level with increasing number of local classifiers, since the LOG-LAT metric features a network wide normalization and thus contains global information.

With increasing number of available individual genotypes, the Business topic requires consistently fewer local classifiers than the Celebrities. One explanation of this might be a higher heterogeneity of sub-topics within Celebrties and hence lower topic-wide response consistency. For example, many businesses and brand names are designed to be topically distinct, while celebrities may be perceived as sports stars, politicians, or company executives. For topics like the latter, more individual genotypes are needed to arrive at a correct hashtag classification.

It is important to note that we use classification only as a way to evaluate if the topic specific-behavior captured by our genotype metrics is invariant for users. While the genotypes might be adopted for actual novel information classification into topics, an improved classifier for such applications may benefit from combining the genotypes with textual features of tweets.

6.4 Discussion

When comparing the results of this chapter to the results of Chapter 5, it becomes more clear why the agent-based network model discussed in Chapter 2 is not able to fully explain the fluctuations that are observed in real hashtag adoption data on the Twitter

social network. In Section 2.2, the coupling strength between agents in the network is assumed to be defined by the relative frequency of interactions between each pair of neighboring agents. However, as shown in Section 6.3, we found that response time (TIME) and number of hashtag uses (N-USES) within a topical network perform worse than the null model (Random Error) and are the least reliable predictors of adoption behavior. This indicates that each user is typically inconsistent in their response times to same-topic hashtags, and suggests that an agent-based ODE model parameterized by time is unlikely to accurately capture adoption behavior on social networks similar to Twitter. Over all, the evidence presented in this chapter supports the results of Chapter 5.

In future work, we are interested in applying the genotype framework beyond hashtags and Twitter. Alternative information retrieval and natural language processing approaches for annotating tweets into topics can also be adopted within our framework. Hashtags, as a means of annotation and defining a universal vocabulary, are also common in systems for other types of content such as music, photos and video. Examples include the photo sharing social site Flickr, the video sharing site YouTube, and music streaming sites such as Last.fm and Pandora. We believe that our hashtag-based genotype framework might extend to modeling and analysis of user behavior when interacting and disseminating photos and multimedia as well.

We adopt a model in which every information item (hashtag) is associated with exactly one topic. This particular way to instantiate our genotype model is the first attempt to

demonstrate the preserved behavior within a topic. One can naturally extend this to a richer analysis in which we have "soft" association of content items and topics. One promising direction is to learn such association using latent topic models such as the ones introduced by Blei and colleagues [110] in lieu of hard topic classification. Our proposed applications (topic prediction, latency minimization, and adoption prediction) can then be extended naturally using the probabilistic association weights of hashtags for different topics.

Chapter 7

Conclusions

Many models of disease and rumor spreading phenomena average the behavior of individuals in a population in order to obtain a coarse description of expected system behavior. For these types of models, we determined how close the coarse approximation is to its corresponding agent-based system. These findings lead to a general result on the logistic behavior of information propagation for networks on both connected graphs with doubly stochastic edge weights, and connected graphs with symmetric edge weights. Moreover, we discussed the appropriateness of the discrete logistic approximation for a few example heterogeneous graph topologies.

Motivated by our need to test the logistic approximation results with real hashtag adoption data from the Twitter social network, we used statistical learning methods to construct an adaptive state estimator for nonlinear systems. Optimal state estimation typically requires knowledge of process and measurement uncertainty, which we proposed can be estimated from (conditioned on) past observed data. As new data is acquired, the state estimates, process uncertainty, and measurement uncertainty were updated accord-

ingly. These statistical estimation methods helped us compare the real Twitter hashtag data to the agent-based model, We found that the agent-based model, where coupling between agents is only described in terms of network structure, does not sufficiently capture the user adoption behavior of hashtags, and thus a more descriptive user model is required.

Since information propagation in social media depends on the topic-specific user behavior, we developed a novel model incorporating dynamic user behavior, termed a *genotype*. The genotype is a *per-topic* summary of a user's interest, activity and susceptibility to adopt new information. We demonstrated that user genotypes remain invariant within a topic by applying the genotypes for the classification of new information spread in large-scale real networks (demonstrated 87% accuracy). There is still room for this genotype framework to be developed, and we leave it as future work to continue in this direction.

Bibliography

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] M. Busch and J. Moehlis, "Homogeneous assumption and the logistic behavior of information propagation," *Phys. Rev. E*, vol. 85, no. 2, p. 026102, 2012, copyright 2014 by the American Physical Society.

[3] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13, 2013, pp. 236–242, doi: 10.1145/2492517.2492621.

[4] P. Bogdanov, M. Busch, J. Moehlis, A. Singh, and B. K. Szymanski, "Modeling individual topic-specific behavior and influence backbone networks in social media," *Social Network Analysis and Mining*, vol. 4, no. 1, 2014.

[5] M. Busch and J. Moehlis, "A nonparametric adaptive nonlinear statistical filter," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, Dec. 2014, (To appear). Copyright 2014 IEEE.

[6] D. Daley and D. Kendall, "Stochastic rumors," *J. Inst. Maths Applics*, vol. 1, pp. 42–55, 1965.

[7] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, "Theory of rumour spreading in complex social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457 – 470, 2007.

[8] Y. Moreno, M. Nekovee, and A. F. Pacheco, "Dynamics of rumor spreading in complex networks," *Phys. Rev. E*, vol. 69, no. 6, p. 066130, 2004.

[9] J. Zhou, Z. Liu, and B. Li, "Influence of network structure on rumor propagation," *Physics Letters A*, vol. 368, no. 6, pp. 458 – 463, 2007.

[10] L. Bettencourt, A. Cintron-Arias, D. Kaiser, and C. Castillo-Chavez, "The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological

models," *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 513 – 536, 2006.

[11] Y. Bulygin, "Epidemics of mobile worms," in *Performance, Computing, and Communications Conference*, 2007, pp. 475 –478.

[12] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Phys. Rev. E*, vol. 66, no. 3, p. 035101, 2002.

[13] L. Allen, "Some discrete-time SI, SIR, and SIS epidemic models," *Mathematical Biosciences*, vol. 124, no. 1, pp. 83 – 105, 1994.

[14] L. Allen and A. Burgin, "Comparison of deterministic and stochastic SIS and SIR models in discrete time," *Mathematical Biosciences*, vol. 163, no. 1, pp. 1 – 33, 2000.

[15] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.

[16] H. W. Hethcote and J. W. V. Ark, "Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs," *Mathematical Biosciences*, vol. 84, no. 1, pp. 85 – 118, 1987.

[17] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[18] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks," *Journal of Theoretical Biology*, vol. 235, no. 2, pp. 275 – 288, 2005.

[19] C. Castellano and R. Pastor-Satorras, "Thresholds for epidemic spreading in networks," *Phys. Rev. Lett.*, vol. 105, no. 21, p. 218701, 2010.

[20] R. Durrett, "Some features of the spread of epidemics and information on a random graph," *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 107, no. 10, pp. 4491–4498, 2010.

[21] J. Lloyd-Smith, S. Schreiber, and W. Getz, "Moving beyond averages: Individual-level variation in disease transmission," *Contemporary Mathematics*, vol. 410, pp. 235–234, 2006.

[22] R. M. May and A. Lloyd, "Infection dynamics on scale-free networks," *Phys. Rev. E*, vol. 64, no. 6, p. 066112, 2001.

[23] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, "Epidemic outbreaks in complex heterogeneous networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 26, pp. 521–529, 2002.

[24] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, 2001.

[25] J. Saramäki and K. Kaski, "Modelling development of epidemics with dynamic small-world networks," *Journal of Theoretical Biology*, vol. 234, no. 3, pp. 413 – 421, 2005.

[26] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, 2002.

[27] M. Morris, "Data driven network models for the spread of infectious disease," in *Epidemic Models: Their Structure and Relation to Data*, D. Mollison, Ed. Cambridge, Great Britain: Cambridge University Press, 1995, pp. 302–322.

[28] A. Lloyd, S. Valeika, and A. Cintron-Arias, "Infection dynamics on small-world networks," *Contemporary Mathematics*, vol. 410, pp. 209–234, 2006.

[29] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Characterization and modeling of weighted networks," *Physica A*, vol. 346, no. 1-2, pp. 34 – 43, 2005.

[30] C. Castellano and R. Pastor-Satorras, "Non-mean-field behavior of the contact process on scale-free networks," *Phys. Rev. Lett.*, vol. 96, no. 3, p. 038701, 2006.

[31] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," *American Journal of Sociology*, vol. 113, no. 3, pp. 702 – 734, 2007.

[32] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, "Discrete-time Markov chain approach to contact-based disease spreading in complex networks," *Europhysics Letters*, vol. 89, no. 3, p. 38009, 2010.

[33] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. pp. 167–256, 2003.

[34] C. Godsil and G. Royle, *Algebraic Graph Theory*, ser. Graduate texts in mathematics. Springer-Verlag, New York, 2001, vol. 207.

[35] P. Sharma, U. Khurana, B. Shneiderman, M. Scharrenbroich, and J. Locke, "Speeding up network layout and centrality measures for social computing goals," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, ser. Lecture Notes in Computer Science, J. Salerno, S. Yang, D. Nau, and S.-K. Chai, Eds. Springer Berlin / Heidelberg, 2011, vol. 6589, pp. 244–251.

[36] T. G. Hallam and C. E. Clark, "Non-autonomous logistic equations as models of populations in a deteriorating environment," *Journal of Theoretical Biology*, vol. 93, no. 2, pp. 303 – 311, 1981.

[37] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks*. Princeton, NJ: Princeton University Press, 2009.

[38] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, "Analysis of a large-scale weighted network of one-to-one human communication," *New J. Phys.*, vol. 9, no. 6, p. 179, 2007.

[39] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, 2009, pp. 497–506.

[40] L. Allen, *An Introduction to Stochastic Processes with Applications to Biology*. Upper Saddle River, NJ: Pearson Education, Inc., 2003.

[41] B. Gharesifard and J. Cortés, "When does a digraph admit a doubly stochastic adjacency matrix?" in *American Control Conference (ACC)*, 2010, pp. 2440 –2445.

[42] R. Ren and R. W. Beard, *Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications*. Springer-Verlag, London, 2008.

[43] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215 –233, 2007.

[44] H. Dym, *Linear Algebra in Action*. Providence, RI: American Mathematical Society, 2007.

[45] S. J. Leon, *Linear Algebra with Applications*. Upper Saddle River, NJ: Prentice Hall, 1998.

[46] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Hanover, MA: Now Publishers Inc., 2006.

[47] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Wikipedia vote network," http://snap.stanford.edu/data/wiki-Vote.html, 2008.

[48] A. L. Edwards, *An Introduction to Linear Regression and Correlation*. W. H. Freeman & Co., 1984.

[49] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, 2011, pp. 177–186.

[50] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10, 2010, pp. 591–600.

[51] O. Tsur and A. Rappoport, "What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12, 2012, pp. 643–652.

[52] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11, 2011, pp. 695–704.

[53] R. Ribiero, "25 small-business Twitter hashtags to follow," http://www.biztechmagazine.com/article/2012/06/25-small-business-twitter-hashtags-follow, 2012.

[54] A. K. McCallum, "MALLET: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.

[55] J. Rennie, "20 newsgroups," http://www.qwone.com/~jason/20Newsgroups/, 2008.

[56] A. Gulli, "News space," http://www.di.unipi.it/~gulli/, 2012.

[57] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential Twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10, 2010, pp. 261–270.

[58] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in Twitter," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12, 2012, pp. 251–260.

[59] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[60] G. Strang, *Introduction to Linear Algebra*. Wellesley, MA: Wellesley-Cambridge Press, 2011.

[61] D. Simon, *Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches*. John Wiley & Sons, 2006.

[62] F. Schlee, C. Standish, and N. Toda, "Divergence in the Kalman filter." *AIAA Journal*, vol. 5, no. 6, pp. 1114–1120, 1967.

[63] R. Fitzgerald, "Divergence of the Kalman filter," *Automatic Control, IEEE Transactions on*, vol. 16, no. 6, pp. 736–747, 1971.

[64] B. M. Åkesson, J. B. Jørgensen, N. K. Poulsen, and S. B. Jørgensen, "A Kalman filter tuning tool for use with model-based process control," 2007.

[65] Y. Oshman and I. Shaviv, "Optimal tuning of a Kalman filter using genetic algorithms," *Sort*, vol. 6, p. 3, 2000.

[66] T. D. Powell, "Automated tuning of an extended Kalman filter using the downhill simplex algorithm," *Journal of Guidance, Control, and Dynamics*, vol. 25, no. 5, pp. 901–908, 2002.

[67] F. Ding, Y. Liu, and B. Bao, "Gradient-based and least-squares-based iterative estimation algorithms for multi-input multi-output systems," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 226, no. 1, pp. 43–55, 2012.

[68] T. K. Lau and K.-w. Lin, "Evolutionary tuning of sigma-point Kalman filters," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 771–776.

[69] Q. Song and J.-D. Han, "An adaptive UKF algorithm for the state and parameter estimations of a mobile robot," *Acta Automatica Sinica*, vol. 34, no. 1, pp. 72 – 79, 2008.

[70] V. Fathabadi, M. Shahbazian, K. Salahshour, and L. Jargani, "Comparison of adaptive Kalman filter methods in state estimation of a nonlinear system using asynchronous measurements," in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2009.

[71] J. R. Forbes, "Adaptive approaches to nonlinear state estimation for mobile robot localization: an experimental comparison," *Transactions of the Institute of Measurement and Control*, vol. 35, no. 8, pp. 971–985, 2013.

[72] C. Hajiyev and H. E. Soken, "Robust adaptive Kalman filter for estimation of UAV dynamics in the presence of sensor/actuator faults," *Aerospace Science and Technology*, vol. 28, no. 1, pp. 376 – 383, 2013.

[73] Z. Jiang, Q. Song, Y. He, and J. Han, "A novel adaptive unscented Kalman filter for nonlinear estimation," in *Decision and Control, 2007 46th IEEE Conference on*, 2007, pp. 4293–4298.

[74] M. Karasalo and X. Hu, "An optimization approach to adaptive Kalman filtering," *Automatica*, vol. 47, no. 8, pp. 1785–1793, 2011.

[75] S. Kosanam and D. Simon, "Kalman filtering for uncertain noise covariances," Ph.D. dissertation, Cleveland State University, 2004.

[76] Y. Li and J. Li, "Robust adaptive Kalman filtering for target tracking with unknown observation noise," in *Control and Decision Conference (CCDC), 2012 24th Chinese*, 2012, pp. 2075–2080.

[77] R. Mehra, "Approaches to adaptive filtering," *Automatic Control, IEEE Transactions on*, vol. 17, no. 5, pp. 693–698, 1972.

[78] S. Sarkka and J. Hartikainen, "Non-linear noise adaptive Kalman filtering via variational Bayes," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, 2013, pp. 1–6.

[79] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *Automatic Control, IEEE Transactions on*, vol. 54, no. 3, pp. 596–600, 2009.

[80] R. G. Miller, "The jackknife-a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.

[81] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, vol. 9, no. 3, pp. 586–596, 1981.

[82] J. Shao, C. J. Wu *et al.*, "A general theory for jackknife variance estimation," *The Annals of Statistics*, vol. 17, no. 3, pp. 1176–1197, 1989.

[83] J. Shao, "Consistency of least-squares estimator and its jackknife variance estimator in nonlinear models," *Canadian Journal of Statistics*, vol. 20, no. 4, pp. 415–428, 1992.

[84] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," *Journal of Geophysical Research: Oceans (1978–2012)*, vol. 99, no. C5, pp. 10 143–10 162, 1994.

[85] ——, "The ensemble Kalman filter: Theoretical formulation and practical implementation," *Ocean Dynamics*, vol. 53, no. 4, pp. 343–367, 2003.

[86] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Courier Dover Publications, 2007.

[87] J. Shao, "The efficiency and consistency of approximations to the jackknife variance estimators," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 114–119, 1989.

[88] G. A. Ghazal and H. Neudecker, "On second-order and fourth-order moments of jointly distributed random matrices: a survey," *Linear Algebra and its Applications*, vol. 321, no. 1, pp. 61–93, 2000.

[89] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[90] D. Higham., "An algorithmic introduction to numerical simulation of stochastic differential equations," *SIAM Review*, vol. 43, no. 3, pp. 525–546, 2001.

[91] A. S. Hurn, K. A. Lindsay, and V. L. Martin, "On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations*," *Journal of Time Series Analysis*, vol. 24, no. 1, pp. 45–63, 2003.

[92] U. Picchini, "Stochastic differential equations toolbox," http://www.mathworks.com/matlabcentral/linkexchange/links/1387-stochastic-differential-equations-toolbox, Jun 2009.

[93] J. Lagarias, J. Reeds, M. Wright, and P. Wright, "Convergence properties of the nelder–mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[94] D. S. Moore, *The Basic Practice of Statistics*. Palgrave Macmillan, 2010.

[95] L. R. Petzold and U. M. Ascher, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia, PA: SIAM, 1998.

[96] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03, 2003, pp. 137–146.

[97] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Extracting influential nodes on a social network for information diffusion," *Data Mining and Knowledge Discovery*, vol. 20, no. 1, pp. 70–97, 2010.

[98] N. Friedkin, *A Structural Theory of Social Influence*. Cambridge University Press, 2006, vol. 13.

[99] P. Dodds, K. Harris, I. Kloumann, C. Bliss, and C. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS One*, vol. 6, no. 12, p. e26752, 2011.

[100] R. Bandari, S. Asur, and B. Huberman, "The pulse of news in social media: Forecasting popularity," in *International AAAI Conference on Weblogs and Social Media*, 2012.

[101] C. Lin, Q. Mei, Y. Jiang, J. Han, and S. Qi, "Inferring the diffusion and evolution of topics in social communities," *Social Network Mining and Analysis*, vol. 3, no. d4, p. d5, 2011.

[102] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 2010, pp. 177–184.

[103] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *International AAAI Conference on Weblogs and Social Media*, 2010.

[104] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13, 2013, pp. 23–32.

[105] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, 2011, pp. 65–74.

[106] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, 2011, pp. 45–54.

[107] F. Reali and T. L. Griffiths, "Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift," *Proc. of the Royal Society B: Biological Sciences*, vol. 277, no. 1680, pp. 429–436, 2010.

[108] M. De Choudhury, "Tie formation on Twitter: Homophily and structure of egocentric networks," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011, pp. 465–470.

[109] W. Krzanowski and W. Krzanowski, *Principles of multivariate analysis*. Oxford University Press, 1996.

[110] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

Appendix A

Logistic Bounds of the Completely Connected Solution

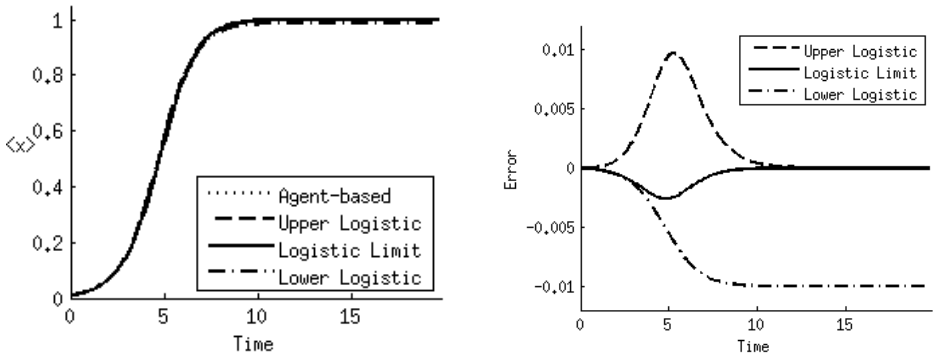
For systems of finite size, it is possible to bound (2.6) by discrete logistic functions. Since the fact that each element of $p_t \in [0,1]$ implies that $\left(p_t^{(i)}\right)^2 \leq p_t^{(i)}$ and thus $|p_t|_2^2 \leq |p_t|_1$, then one obtains the upper bound:

$$x_{t+h} \leq x_t + h\beta_t \frac{N}{N-1} x_t \left(1 - x_t\right). \tag{A.1}$$

A lower bound for (2.6) can be obtained by simply truncating the the $|p_t|_2^2$ term:

$$x_{t+h} \geq x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right). \tag{A.2}$$

We now compare the various logistic approximations to conclude that the single step upper and lower logistic bounds of (2.6) produce upper and lower solutions for all time steps. First, we show that if given $x_t, y_t \in [0,1]$ at time t and parameters $\phi, \theta \in R_{>0}$, then $x_t \geq y_t$ implies $x_{t+h} \geq y_{t+h}$ if the two points evolve according to the same discrete logistic equation of the form $x_{t+h} = x_t + \phi h \beta_t x_t \left(1 - \theta x_t\right)$.



(a) (b)

Figure A.1: (a) Comparison of the graph-based solution to the logistic equation in the thermodynamic limit, as well as the upper and lower bounding logistic solutions for the finite case. The solutions are nearly indistinguishable. (b) Pointwise error difference of the upper, lower, and thermodynamic limit logistic solutions with respect to the graph-based solution. Parameters are $\beta_t = 1$ and $N = 100$ in both plots.

By assuming $x_t \geq y_t$, then

$$\begin{aligned} x_{t+h} - y_{t+h} &= x_t + \phi h \beta_t x_t \left(1 - \theta x_t\right) - y_t - \phi h \beta_t y_t \left(1 - \theta y_t\right) \\ &= \left(x_t - y_t\right) + \phi h \beta_t \left(\left(x_t - y_t\right) - \theta \left(x_t^2 - y_t^2\right)\right) \\ &= \left(x_t - y_t\right) \left[1 + \phi h \beta_t \left(1 - \theta \left(x_t + y_t\right)\right)\right] \\ &\geq \left(x_t - y_t\right) \left[1 + \phi h \beta_t \left(1 - 2\theta\right)\right], \end{aligned}$$

and is non-negative when $2\theta \leq 1$ or

$$h \leq \frac{1}{\sup_t \beta_t \phi \left(2\theta - 1\right)}. \tag{A.3}$$

By imposing a requirement on the step size of the discrete logistic equation, it can be shown that one discrete logistic solution bounds another discrete logistic solution if their single step dynamics also bound each other. When applied directly to (A.1) and (A.2), the minimum step size comes from (A.2) with $\phi = 1$ and $\theta = N/\left(N - 1\right)$.

Now suppose h satisfies (A.3) with the ϕ and θ values from (A.2) so that

$$h \leq \frac{N-1}{N+1}\frac{1}{\sup_t \beta_t}.$$

The dynamic equation (2.7) yields a logistic approximation, x_t^*, that is bounded by (A.1) and (A.2). To show this is true, let

$$\begin{aligned} x_{t+h}^{Upper} &:= x_t + h\beta_t \frac{N}{N-1} x_t \left(1 - x_t\right), \\ x_{t+h}^{*} &:= x_t + h\beta_t x_t \left(1 - x_t\right), \\ x_{t+h}^{Lower} &:= x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right). \end{aligned}$$

Given an initial condition $|P_0|_1 / N = x_0^{Upper} = x_0^* = x_0^{Lower}$, equations (2.7) - (A.2) provide the respective x_1 values and their relation to each other (i.e., $x_1^{Lower} \leq x_1^* \leq x_1^{Upper}$). Let the relationship between x_0 and the x_1 terms be the base case for induction.

To show that $x_t^{Upper} \geq x_t^*$ implies $x_{t+h}^{Upper} \geq x_{t+h}^*$ as the inductive step, we begin with the hypothesis that $x_t^{Lower} \leq x_t^* \leq x_t^{Upper}$. If one were to briefly let $y_t^{Lower} = x_t^*$ and $y_t^{Upper} = x_t^*$, then the same procedure used to obtain the base cases asserts that $y_{t+h}^{Lower} \leq x_{t+h}^* \leq y_{t+h}^{Upper}$. Since $x_t^{Lower} \leq y_t^{Lower}$ and $y_t^{Upper} \leq x_t^{Upper}$, it follows that $x_{t+h}^{Lower} \leq y_{t+h}^{Lower}$ and $y_{t+h}^{Upper} \leq x_{t+h}^{Upper}$. Hence, $x_t^{Lower} \leq x_t^* \leq x_t^{Upper}$ implies $x_{t+h}^{Lower} \leq x_{t+h}^* \leq x_{t+h}^{Upper}$ so that the solution to (2.7) is bounded by the solutions to (A.1) and (A.2).

For this bounding statement to be true, the step size must be chosen appropriately based on the size on the network and the transmission rate. Thus, when comparing data to the model under the homogeneous assumption, one must consider the behavior of the

information since it affects the transmission rate. The transmission rate affects the step size, which in turn affects the adjacency matrix (in the more general case).

Appendix B

Proof of General Coarse Approximation

B.1 Doubly Stochastic Matrices

To show how well the scalar logistic model approximates the graph-based model, first let A be the doubly stochastic and irreducible adjacency matrix that corresponds with the network topology. Since the elements of p_t are all nonnegative, the one norm of the vector p_t is simply a sum over all of its elements (i.e. $|p_t|_1 = 1_n^T p_t$), we begin by taking the 1-norm of (2.4):

$$\begin{aligned} |p_{t+h}|_1 &= 1^T \left(p_t + h\beta_t A p_t - h\beta_t diag\{p_t\} A p_t\right) \\ &= 1_n^T p_t + h\beta_t 1^T A p_t - h\beta_t p_t^T A p_t. \end{aligned}$$

Quadratic forms have the property that $p_t^T A p_t = p_t^T \left(A_S\right) p_t$, where $A_S = \left(A + A^T\right)/2$ is a symmetric matrix. Using the series representation $A_S = \sum_{i=1}^{n} \lambda_i w_i w_i^T$:

$$\begin{aligned} \left|p_{t+h}\right|_1 &= 1_n^T p_t + h\beta_t 1_n^T A p_t - h\beta_t p_t^T \sum_{i=1}^{n} \lambda_i w_i w_i^T p_t \\ &= \left|p_t\right|_1 + h\beta_t 1_n^T A p_t - h\beta_t \lambda_1 p_t^T w_1 w_1^T p_t - h\beta_t p_t^T \sum_{i=2}^{n} \lambda_i w_i w_i^T p_t. \end{aligned} \tag{B.1}$$

Since A is doubly stochastic, the columns of A each sum to 1 so that the second term of (B.1) simplifies to $h\beta_t \left|p_t\right|_1$.

The matrix A_S will also be doubly stochastic, and thus row stochastic. From the Perron-Frobenius theorem [37], $\lambda_1 = 1$ and $w_1 = 1_n/\sqrt{n}$, and denoting the inner product as $\langle \cdot, \cdot \rangle$, the third term of (B.1) can be simplified as follows:

$$-h\beta_t \lambda_1 p_t^T w_1 w_1^T p_t = -h\beta_t \left\langle 1_n/\sqrt{n}, p_t \right\rangle^2 = -h\beta_t \frac{1}{n} \left\langle 1, p_t \right\rangle^2 = -h\beta_t \frac{1}{n} \left|p_t\right|_1^2.$$

By applying the inner product notation to the fourth term of (B.1), it can be rewritten as $-h\beta_t \sum_{i=2}^{n} \lambda_i \left\langle w_i, p_t \right\rangle^2$.

Upper and lower bounds on the fourth term of (B.1) can be obtained by observing that

$$-\left|\lambda_2\right| \sum_{i=2}^{n} \left\langle w_i, p_t \right\rangle^2 \leq \sum_{i=2}^{n} \lambda_i \left\langle w_i, p_t \right\rangle^2 \leq \left|\lambda_2\right| \sum_{i=2}^{n} \left\langle w_i, p_t \right\rangle^2.$$

To further simplify this expression, we can use the submultiplicative property of matrix norms, where $\sum_{i=2}^{n} \left\langle w_i, p_t \right\rangle^2 \leq \sum_{i=1}^{n} \left\langle w_i, p_t \right\rangle^2 = \left\|W^T p_t\right\|_2^2 \leq \left\|W^T\right\|_2^2 \left|p_t\right|_2^2 = \left|p_t\right|_2^2$. By substituting $\sigma_2 = \left|\lambda_2\right|$ and applying this inequality, one obtains the following:

$$-\sigma_2 \left|p_t\right|_2^2 \leq \sum_{i=2}^{n} \lambda_i \left\langle w_i, p_t \right\rangle^2 \leq \sigma_2 \left|p_t\right|_2^2,$$

which indicates that the fourth term of (B.1) is a term of order $h\sigma_2$.

Therefore, (B.1) simplifies to

$$|p_{t+h}|_1 = |p_t|_1 + h\beta_t |p_t|_1 - \frac{h\beta_t}{n} |p_t|_1^2 + O(h\sigma_2). \tag{B.2}$$

Finally, divide by n, and let $x_t = |p_t|_1 / n$ to obtain the average probability of being informed:

$$x_{t+h} = x_t + h\beta_t x_t (1 - x_t) + O(h\sigma_2). \tag{B.3}$$

B.2 Symmetric Matrices

Beginning with expression (2.14), one can factor out R_1 use the fact that $(A - R_1)$ is a symmetric matrix to obtain

$$p_{t+h} = p_t + h\beta_t (I - diag\{p_t\}) p_t + h\beta_t (I - diag\{p_t\}) WDW^T p_t,$$

where D is a matrix whose diagonal elements are the eigenvalues of $(A - R_1)$. By recognizing that $-\sigma_1 I \leq WDW^T \leq \sigma_1 I$, it follows that one obtains (2.15) by summing the elements and dividing by the cardinality of the population. A similar argument shows that mean-field solutions are upper-bounded by solutions to

$$x_{t+h} = x_t + \sigma_1 h\beta_t x_t (1 - x_t),$$

where $x_t = |p_t|_1 / N$ and σ_1 provides a time-scaling effect on the step-size when h satisfies (A.3).

B.3 Row Stochastic Upper Bound

Suppose that the network adjacency matrix A is row stochastic so that $A1_N = 1_N$. Beginning with equation (2.15), one is able to factor A out of the expression $A - R_1$ by using the facts that $A1_N = 1_N$ and $R_1 = 1_N 1_N^T/N$ to obtain

$$\begin{aligned} R_1 &= \frac{1}{N} 1_N 1_N^T \\ &= \frac{1}{N} \left(A1_N\right) 1_N^T \\ &= AR_1, \end{aligned}$$

and thus

$$\begin{aligned} \|A - R_1\|_2 &= \|A(I - R_1)\|_2 \\ &\leq \|A\|_2 \|I - R_1\|_2. \end{aligned} \tag{B.4}$$

The matrix $(I - R_1)$ is a circulant and symmetric Toeplitz matrix of dimension N with eigenvalues

$$\lambda_k = 1 + \sum_{k=0}^{N-1} \left(-\frac{1}{N}\right) \exp\left(\frac{i2\pi k}{N}\right). \tag{B.5}$$

For $k = 0$, $\lambda_0 = 1 - (N)(N^{-1}) = 0$. For $k \neq 0$, we use the fact that (B.5) contains a geometric series to obtain

$$\begin{aligned} \lambda_{k\neq 0} &= 1 - \frac{1}{N} \frac{1 - \left(\exp\left(\frac{i2\pi}{N}\right)\right)^N}{1 - \exp\left(\frac{i2\pi}{N}\right)} \\ &= 1. \end{aligned}$$

Therefore, since $\|I - R_1\|_2 = 1$ we obtain from (B.4) the bound $\|A - R_1\|_2 \leq \|A\|_2$. When including the time step h and rate parameter β, one finds that the approximation error

of (2.15) has the bound

$$O(h\|A - R_1\|_2) \leq O(h\beta\|A\|_2),$$

for any row stochastic matrix A.

Appendix C

Additional Definitions

As explained in [1], the characteristic path length (L) of a network is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. Similarly, one can define the characteristic path length (L_i) of a single node i as the average shortest path length between i and each other $j \in V$.

To define a network's average clustering coefficient (C), we first define the set of edges E_i that exist between a given node $i \in V$ and its neighbors as $E_i = \{(i, j) \in E,\ \forall\, j \in V\}$. If node i has k_i neighbors, then the clustering coefficient (C_i) of node i is

$$C_i = \frac{2\,|E_i|}{k_i\,(k_i - 1)}, \tag{C.1}$$

where $|E_i|$ represents the cardinality of E_i, and $k_i\,(k_i - 1)\,/2$ is the maximum number of edges that can possibly exist in E_i. Hence, C is the average value of C_i over all i.

Appendix D

Topic Hashtag Lists

Note: The hashtags "glennbeck", "obama", and "palin" were each verified to belong to both the Celebrity and Politics topics, and "nascar" was verified to belong to both the Politics and Sports topics. For the purposes of this manuscript, these hashtags were treated as distinct elements in each topic (i.e., "obama" referred to as a celebrity versus "obama" referred to as a politician), and with identical genotype metric values. For example, since only the hashtag "obama" is detected in the data, it is understood that references to obama as a celebrity co-occur with references to obama as a politician.

D.1 Business

4jobs, business, collaboration, consumers, ecommerce, economy, entrepreneurs, innovation, leadership, management, marketing, mktg, networking, painatthepump, restaurant, retail, sales, shoplocal, smallbiz, smallbusiness, smallbusinesssaturday, smallsizSat, socbiz, socialbiz, socialbusiness, startups, tax

D.2 Celebrity

aaliyah, anoopdesai, argentinawantsjb, ashleytisdale, australiawantsjonas, brazilmissesdemi, brazilwantsjbagain, brazilwantsjb, bringbackrachelle, bsb, chamillionaire, craigferguson, davidarchuletta, gagavmas, glennbeck, happybirthdaypink, iwantpeterfacinelli, jonaslive, michaeljackson, mileycomeback, mj, niley, obama, palin, regis, signmattgiraud, teamtaylor, tilatequila, welovekevinjonas, weloveyoujoejonas, weloveyoujustin, weloveyoumiley

D.3 Politics

1u, 2012gop, 2012, 250gas, 2nd, 2, 4all2c, 912, 99percent, a4a, abc, abortion, abortions', ac360, acon, agenda21, ak, alabama, algop, alinsky, allenwest, allstar, al, alprimary, america, ampat, ampats, andrewbreitbart, anybodybutobama, armyofbreitbarts, askthe, attackwatch, awesome, axelrod4romney, az, azright, bbc, beck, bell, bet, bettymccollum, bible, biggovernment, bighollywood, bigjournalism, bigot, bigpeace, blacknews, black, blogconclt, boston, boycotthollywood, b, breastlift, breitbartarmy, breitbartishere, breitbartnet, breitbart, breitbart's, brtt, budget, c4l, cain, ca, caring, catcot, catholic, cbsnews, cbs, cfsa, chicago, chitpp, christian, christians, clcs, cnndebate, cnn, college, communism, communist, communists, compromise, congress, con, conservative, conservatives, consnc, constitution, cpac12, cpac, criticalrace, criticalracetheory, cspj, ct, dads, daretovoterick, dc, democrat, democrats, dem, dems, de, dianesawyer, dinnerwithbarack,

dnc2012, dnc, doj, dprs, drudge, edchat, edshow, education, egypt, election2012, election, electionsmatter, endorsemitt, energy, epicfail, espn, exposetheleft, fail, faith, fastandfurious, fbi, ff, film, flgop, florida, fl, flprimary, flsen, fluke, forward, foxnews, fox, fraud, freechrisloesch, freechris, freedom, gamechange, ga, gaprimary, gas, gbtv, gen44, gingrich, glennbeck, goa, god, gop2012, go, gop, g, green, gsa, guns, gu, handsoff, hannity, hbo, hcr, healthcare, hhrs, hi, hispanic, holder, hollywood, ho, humor, hypocrisy, hypocrites, iamandrewbreitbart, iambreitbart, iamthe53, iamthemob, icon, id, illinois, il, imab, imbreitbart, impeach, independent, independents, inde, iranelection, iran, iranrevolution, islam, israel, isreal, isupportrush, jcot, jemuhgreen, jesus, jewish, jews, jihad, jobs, keystone, ks, ktvd, kulaktv, kxl4jobs, la, latino, launfd, liberal, liberals, libertarian, liberty, libya, limbaugh, lnn, lnyhbt, lol, lolwut, l, lur, maddow, majority, ma, mapoli, maprimary, maraliasson, marines, marklevinshow, marxist, may, mdayton, md, media, military, mil, mi, mitt2012, mitt, mmfa, mn2010, mngop, mnleg, moms, mo, mosque, msm, msnbc, ms, muslim, nascar, nationaldebt, navy, nbc, ncgop, nc, ndaa, ndcaucus, nd, news, newt2012hq, newt2012, newt, newyork, newyorkpost, nj, nobama2012, nobama, nolabels, notgoingaway, notobama, n, npr, nra, nugent, nwo, nyc, ny, obama2012, obamaateadog, obamacare, obamadogrecipes, obamafail, obamaland, obamaonempty, obama, obama's, ocares, occupiers, occupy, occupysf, occupyunmasked, occupywallstreet, ocra, ohio, oh, ohprimary, ohyeah, oil, ok, okprimary, omg, o, orcot, orpol, oversight, ows, p21, p2, palin, pa, parents, patriot, patriots, paul, pbs, phnm, pinkslimemedia, plannedparenthood, politics, potus, p, progressive, prolife, propaganda, pr, pushbackgop, r3volution, racecard,

racewar, racism, racistcrt, racistderrickbell, racist, reagan, redeye, religion, repealandreplace, resist44, retweet, right, rino, ri, rnc, romney, ronpaul2012, ronpaul, rosen, rs, rt, rush, sallykohn, sanford, santorum, sarahpalin, savage, saveamerica, sayfie, scgop, scotus, sc, scprimary, seiu, sgp, shariah, sharia, sharpton, socialism, soledad, solyndra, sot, sotu, 's, s, standwithbreitbart, stopmitt, stoprush, stoptweetingsoledad, stribpol, supertuesday, syria, tbrs, tco, tcot, tcot_talk, tc, teambreitbart, teamdueprocess, teamwc, teapa, teapar, teapart, teaparty, tea, teap, te, terrorists, texas, theblaze, thefive, thevetting, tif, timetochoose, timetochoos, tiot, tlot, tn, tnprimary, topprog, t, tpot, tppatriots, tp, tpp, trayvonmartin, trayvon, treasonousacts, treason, truthiness, truth, truthteam, tsa, tsot, tummytuck, twcot, tweetcongress, twiste, twisters, tw, tx, tyranny, ucot, ujcp, undefeated, union, unions, un, usa, usmc, utpol, ve, veteran, veterans, vets, vetthemedia, vetthepress, vettheprez, vi, voteobamaout, vote, voterfraud, voteridnow, voterid, v, vt, vtprimary, wakeup, wa, waronmoms, waronwomen, war, wecantwait, weeklyrecap, wethepeople, wethepeo, whitehouse, whyimin, winning, wi, wiprimary, wirecall, wiright, wisconsin, withnewt, wiunion, woman, women4newt, women, wow, w, wv, wwiimuseum, wy, yal, zerobama, zimmerman

D.4 Science and Technology

140conf, advertising, amazon, android, apple, apps, beatcancer, blackberry, books, consciousness, design, digg, digital, drivehertz, drupal, e3, ebay, ecademy, epharma, eventprofs, facebook, fb, firefox, flickr, formspringme, foursquare, free, funny, google,

google+, googlewave, harmony, hcsm, hootsuite, infographic, in, instagram, internet, ipad, iphone, jquery, linkedin, linux, mac, mashable, mhealth, microsoft, mobile, moonfruit, mp3, nokia, openwebawards, peace, photoshop, php, pinterest, pipa, prsa, redessociales, runkeeper, seo, shared, sharepoint, shazam, smartmeters, smm, sm, smtalk, socialmedia, social, socpharm, sopa, squarespace, stopbullying, sundayblessings, sxsw, tchat, teamfollowback, technion, technology, tech, tinychat, trackle, travel, trb, tweetphoto, twibbon, twittermarketing, twitter, unity, wave, webdesign, weworkin, wisdom, wordpress, wwdc, youtubefail, youtube

D.5 Sports

ashes, canucks, comedy, cowboys, cricket, cubs, dodgers, f1, follow, football, golf, goroaddogs, lakers, mets, mlb, mma, nascar, nba, nfl, nhl, phillies, redsox, rio2016, rugby, soccer, sport, sports, tdf, teamzucker, ufc, victorysessions, warriors, yankees