Keeping Rumors in Proportion: Managing Uncertainty in Rumor Systems

Peter M. Krafft
Information School
University of Washington
pmkrafft@uw.edu

Emma S. Spiro
Information School
University of Washington
espiro@uw.edu

ABSTRACT

The study of rumors has garnered wider attention as regulators and researchers turn towards problems of misinformation on social media. One goal has been to discover and implement mechanisms that promote healthy information ecosystems. Classically defined as regarding ambiguous situations, rumors pose the unique difficulty of intrinsic uncertainty around their veracity. Further complicating matters, rumors can serve the public when they do spread valuable true information. To address these challenges, we develop an approach that reifies "rumor proportions" as central to the theory of systems for managing rumors. We use this lens to advocate for systems that, rather than aiming to stifle rumors entirely or aiming to stop only false rumors, aim to prevent rumors from growing out of proportion relative to normative benchmark representations of intrinsic uncertainty.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → *Psychology*; *Sociology*; • **General and reference** → Design.

KEYWORDS

Rumors; design; misinformation; uncertainty

ACM Reference Format:
Peter M. Krafft and Emma S. Spiro. 2019. Keeping Rumors in Proportion: Managing Uncertainty in Rumor Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300876

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300876

1 INTRODUCTION

For the past several years the city of Seattle, Washington has been in the midst of a housing crisis. The crisis is popularly attributed to the influx of tech workers driving prices up, leading to both housing development and displacement. In the midst of this crisis, discussion and rumors spread about whether rising prices might actually reflect a bubble. Recently several new facts have come to light. Stories swirl around the city bars—empty units in new buildings, rents beginning to deflate, the Seattle-based employer Amazon building new headquarters in other regions. Given these facts about the situation, a gambler might place certain odds on whether there is in fact a housing bubble. But does the level of public concern reflect those rationally calculated odds? What if the current of vague worries transforms into a wave of intense panic?

It is impossible to tell whether people who insist there is an impending housing crash are right or wrong, but if these voices become loud enough, perhaps disproportionately loud, there could be material consequences. At the same time, if material consequences are inevitable, if for instance the real estate industry has simply been maneuvering to prop up prices, then people who believe prices will continue to increase might make misinformed major financial decisions. So when the latest article in the Seattle Times on the housing crisis gets published, how should this news organization moderate the discussion section on the article's webpage? What does a healthy conversation on this sensitive, contested issue look like? What can be done to manage the uncertainty around these rumors—to preserve discussion but prevent rumors from growing out of proportion?

The study of rumors has recently garnered wider attention as regulators and researchers turn their eyes towards the problems of misinformation and disinformation on social media. The shared hope is that there may be mechanisms we can discover and implement to ensure healthy information ecosystems online. Early in 2018 in his Congressional hearing on the impact of Facebook on the 2016 U.S. presidential election, Mark Zuckerberg stated: "We need to make sure that people aren't using [the voice we give them] ... to spread misinformation." Mr. Zuckerberg's perspective reflects a common tack in the literature on systems for dealing

with problematic forms of information—detect and stop the spread of problematic content [53].

Misinformation is defined as an unintentionally false utterance, while disinformation is information that is deliberately false or misleading [24]. Rumors include a broader class of "unverified information" [7], which can be true or false, and therefore can become disinformation or misinformation once the veracity is settled. Sometimes rumors are simple gossip, perhaps at most irksome for some involved. Other times, rumors are powerful. In times of crisis, rumors draw their strength from the anxiety of uncertainty [1, 3, 36–38]. False rumors can have profound, and unwarranted, negative impact [1, 8, 49]. In these cases false rumors can feel arbitrary, unpredictable, and threatening.

Because of these factors, the focus of much existing work has been to quell false rumors. While valuable for certain purposes, this existing work and rhetoric belies one of the key difficulties of rumors as a broad category. Rumor, as unverified information, includes as a direct consequence a degree of uncertainty. A rumor, in its classical scholarly definition, is only a rumor if its participants cannot tell whether it is true or false. Rumors only exist in ambiguous situations. The fundamental problem with a detect-and-stop approach is then this: How can we be expected to classify a rumor as true or false at the time of spreading if the rumor, by definition, pertains to a proposition for which there is no available definitive evidence to establish truth or falsity? And should we put ourselves in the business of stemming rumors that might turn out to be true?

When ambiguity is intrinsic to the state of the world, we can at best hope to quantify rumor uncertainty given the information available. Although much existing work recognizes this difficulty to some degree, it is often operationalized away in the relevant literature within the human-computer interaction, computer-supported cooperative work, and natural language processing communities. Pieces of work on rumors that neglect uncertainty in their analysis either focus on analyzing rumors in retrospect, after they have been confirmed or debunked, or utilize a broader definition of rumor to include verifiable misinformation [17, 42, 50]. We cannot rely on this common approach for all problems of rumor. Efforts to detect and stop verifiable misinformation or verifiably false rumors are laudable. Techniques should undoubtedly be developed to identify and prevent disinformation campaigns, abusive or hateful behavior, and other clearly problematic content. In addition to these efforts, though, we must also take ambiguous rumors seriously. Here we contribute to the investigation of how to manage uncertainty on digital platforms, how to promote healthy sensemaking of complicated situations, and how to keep rumors in proportion when we don't know what is true.

2 EXISTING APPROACHES

Our key contribution is to lay out a framework and formulate design targets that incorporate considerations of intrinsic uncertainty in systems for managing rumors. The rhetorical and practical focus in much existing work has been on false rumors, and especially verifiably false rumors. We have already outlined our argument for why the goal of preventing only and all false rumors is untenable—an important category of rumors occurs in situations where the veracity of the rumor cannot be determined by the information available about the situation. Before we introduce our own framework, we first discuss limitations of existing approaches to dealing with such cases.

Promoting Critical Thinking

One reasonable approach to take in system design, education, and policy-making is to promote critical thinking. An example of this approach relevant to both disinformation and rumors is educational campaigns such as the popular information literacy course, *Calling Bullshit*.[1] If it is not possible to tell what is true with certainty, perhaps we can foster rational degrees of skepticism in consumers of content and thereby promote self-moderation in rumoring.

This attractive goal seems achievable in principle, and solutions oriented towards it may indeed yield productive outcomes, but the science on belief formation suggests that we will need more. Results in psychology and cognitive science have suggested that people have a tendency to think in ways that are contrary to rational internal representations of uncertainty about facts and situations. One facet of these findings is our tendency to hold relatively "binary" beliefs—beliefs that the world is simply one way or another—rather than nuanced judgments such as based on finer-grained probability assessments [47]. Belief change can occur in this model, but it happens in discrete bursts from disbelief to belief rather than through tempered, gradual changes in shades [6]. Furthermore, instead of always believing the proposition that is most likely according to available evidence, people across a range of laboratory experiments demonstrate behavior consistent with randomly choosing between discrete beliefs with probability related to the evidence [25, 39, 47, 48]. Taking this cognitive science at face value we can conclude that human belief formation displays a degree of arbitrariness that is contrary to the mental representation of rational uncertainty or even to the rational selection of discrete or binary beliefs. To put it simply, people can be simultaneously opinionated, stubborn, and capricious in what they believe. To manage rumors as effectively as we can, we will need to do more than just campaign for critical thinking.

[1] https://www.callingbullshit.org/

Preventing Rumoring

Another reasonable approach, one that circumvents rather than directly addresses the issue of uncertainty, is to try to prevent all rumors regardless of their veracity. This approach is advocated by prominent researchers of rumor in psychology and organizational behavior [14, 15], and has also been explored in computer science (e.g., [46]). Evidence suggests that an indirect approach of ameliorating anxiety about issues that would stimulate rumors can prevent rumors from forming in the first place [14, 15]. Recent studies have indicated that detecting rumor-related content using discussion patterns and content characteristics can also be successful [32, 34, 35, 52, 53].

A disadvantage of the goal of preventing all rumors regardless of veracity is that success at this goal undermines the positive effects that rumoring and true rumors can have. An alternative definition that has been proposed for rumors is "informal news" [40]. This notion highlights that rumors concern topics that are important to people but which are difficult for traditional news outlets to report on because of the ambiguity that surrounds them. Relatedly, one important theory of rumors is that rumors function to help us make sense of the world together [40]. In this capacity and also as a means of expression, rumors also are theorized to serve to reduce our anxieties [43]. Although rumors could be an ineffective mechanism for this function if they are often untrue, both classic and recent studies that have conducted open-ended surveys of rumors in circulation have concluded that most rumors are true or are based in truth [8, 54]. If we stamp out all uncertain rumors, then we will in many cases lose the flag-bearers of beliefs that turn out to be true. Finally, there has historically be a political dimension to rumor, with rumor being a lever of power not just for the political elite but also for the masses [12]. While the approach of preventing all rumors regardless of veracity would certainly have upsides, to stem all rumors would also be to prevent an outlet for informal news, to restrict our ability for collective sensemaking about the world, to undercut expressions of anxiety, and to remove an axis of political power from the disenfranchised. A certain amount of rumor is not just unavoidable but can actually be advisable, potentially even an integral part of a thriving information ecosystem.

Fact-Checking

A final common approach to dealing with rumors is fact-checking. A typical process of fact-checking involves collecting curated public evidence or investigating to find new evidence to determine whether a rumor is true or false. Prominent fact checking organizations include Snopes and PolitiFact, but there has also been work in the academic community pushing this approach in new ways, such as predicting credibility through rumor features (e.g., [9, 10, 29]) or through trust networks (e.g., [51]). As with the "detect-and-stop" approach in general, though, fact-checking is inherently limited in situations of intrinsic uncertainty, when there are not enough facts available to check.

3 RUMOR PROPORTIONS

As we have just discussed, existing approaches to managing rumors in situations of intrinsic uncertainty are variously burdened by theoretical, practical, and ethical issues. To stimulate further conversation around new approaches that may overcome some of these limitations, we advocate for a turn towards focusing on analyzing rumor proportions. If we cannot be rid of all false rumors without also stemming some true rumors, if people cannot always be expected to maintain rational representations of uncertainty individually, and if rumors are sometimes beneficial, then we must come to terms with rumors as a part of life and we must ask: When does a rumor become a problem that warrants action? At what point does the number of people who believe a rumor become a sign of a dysfunctional sociotechnical system? And what is a healthy rumor proportion?

We define the proportion of a rumor to be the fraction of people who would express more belief than disbelief in the rumor at a particular point in time. Plenty of existing work has examined rumor proportions. The statistical techniques at our disposal as researchers necessitate that quantitative analyses use summary statistics such as averages, and therefore metrics like rumor proportions. For instance, works studying the effectiveness of interventions on rumors test whether increases or decreases in rumor proportions are achieved (e.g., [17, 22]). Descriptive research has shown that the proportions of affirmations and denials of rumors are related to the probability of the rumors being true (e.g., [32]). Surveys of public opinion invariably report the proportion of people who hold a particular belief, and there are well-known notable cases of public opinion diverging from, for example, the aggregate opinions of experts, such as in beliefs about climate change [2, 13, 23]. Other research shows that certain rumor proportions, such as beliefs about who perpetrated the 9/11 World Trade Center plane attack, can vary widely across countries [44].

Despite this work, there remains a lack of a clear articulation of rumor proportions as a first-class concept central to the theory of systems to manage rumors, rather than as a construct of public opinion simply necessitated by the use of quantitative methods. We argue that rumor proportions can be more than a descriptive gauge of public opinion. **The**

key theoretical angle that we will develop is to view rumor proportions as a population-level representation of intrinsic uncertainty in the world. We can then ask whether rumor proportions correspond to a "rational" level of intrinsic uncertainty about a rumor, rather than asking whether a rumor itself is true or false. In other words, this theoretical lens allows us to formulate a normative target for rumor proportions that incorporates considerations of intrinsic uncertainty, providing a new way to manage rumors whose veracity cannot be determined.

To illustrate what it means to have such a normative benchmark, we can consider how this idea fits next to existing empirical usages of rumor proportions. For instance, Friggeri et al. [17] look at whether a Facebook post getting "snoped" in its comments affects its reshare rate. When a person links to a Snopes article in a comment on a post, that action could potentially increase or decrease the reshare rate of the original post. Of course, if the original post is verifiably true, we might be fine with the comment increasing the reshare rate, and if the original post is verifiably false, we might hope for the comment to decrease the reshare rate. However, the perspective we articulate here makes clear that we must be more careful in our normative judgments about changes in the reshare rate in intermediate, ambiguous cases of intrinsic uncertainty. In such cases, increases or decreases in rumor proportions are only meaningful if they yield final rumor proportions that are above or below or close to some reasonable threshold.

If just a few people believe a rumor, as long as that rumor is not evidently false and potentially dangerous, we might not declare a problem. We can hardly blame individual people for holding beliefs that are within the realm of possibility, beliefs that have some degree of credibility, however slim. If lots of people believe something that is unlikely to be true, then maybe there is a problem. But again, how many is "lots"? When can we say a rumor has grown out of proportion? To foreshadow our answer: If an educated guess might place the odds of a rumor at 50:50, then to return to our example above, if 20% of people believe the rumor, an increase in reshare rate could be considered acceptable; while if 80% of people believe it, we might rather see a decrease in reshare rate—in either case bringing the rumor proportion closer to 50%. Although indirect diagnostics are possible for certain cases, as we will see later, we choose to first elaborate a direct approach of this sort to answering the challenges at hand.

4 "REASONABLE" RUMOR PROPORTIONS

We aim to specify *exactly* at which point a rumor becomes "out of proportion" with available evidence for the rumor. To achieve this level of precision, we must use a mathematical approach. Our overall strategy is to mathematically formulate reasonable normative benchmarks for rumor proportions phrased as design criteria. We will motivate the design criteria we choose both with results from cognitive science and with results from mathematical decision theory. Mathematical approaches are inherently reductive, and our account in this paper will omit many important considerations. The payoffs will be (1) design targets that are specified precisely enough to be used for mathematical optimization in system design, and (2) a connection to mathematical models of belief formation, which allows us to perform an initial assessment of the feasibility of our design targets.

Design Criteria

We formulate our design criteria with respect to a rumor, denoted R. Consistent with the definition of rumor as a proposition for belief [37], we model R as a being a logical proposition about the world. Our framework is intended to pertain to rumors of "fact", as opposed to, for example, political interpretations, ideologically-minded statements, or highly emotional topics. We consider rumors of fact to be those rumors for which a well-informed domain expert could reasonably be expected to provide an authoritative probability judgment. We also restrict our focus to rumors of sufficiently large public interest since the proportions and dynamics of smaller-scale rumors may operate in substantially different ways. Taking the Seattle Housing Crisis case from our introduction, we will use as an example R the proposition "There is a housing bubble in Seattle that will soon pop" for the purposes of illustration of our formalism.

We let N_R be the number of people who are discussing rumor R on a hypothetical social media platform at a particular point in time. We let n_R be the number of people who affirm a positive belief in rumor R. The rumor proportion of R is then $p_R = \frac{n_R}{N_R}$. In light of the discussion in previous sections we advocate for three design criteria in regard to rumor proportions. These design criteria straddle the line between descriptive hypotheses about human behavior and normative targets for human behavior. The criteria specify, in mathematical terms, how we might hope that people behave within a particular rumoring context. Whether or not people do behave in this manner depends on both the inherent tendencies people have and the structure of the system in which they are interacting. Human behavior is always shaped by the institutions surrounding us. The institutional forms or, in the case of rumors the digital platforms that structure our choices and interactions, create rules, affordances, constraints, and incentives for behavior. All of these factors encourage certain behaviors and discourage others. At the same time, we might say there is a "human nature" in the sense that regardless of these differences across contexts, there are certain ways of behaving that people are fundamentally unlikely to display or incapable of displaying. We

attempt to formulate design goals that are achievable given plausible tendencies of human behavior, and we will argue for the feasibility of achieving these criteria. Our design criteria are:

(1) If rumor proportion p_R is at a "reasonable" level, then new evidence for rumor R should increase p_R.
(2) If rumor proportion p_R is at a "reasonable" level, then new evidence against rumor R should decrease p_R.
(3) If rumor R_1 is more plausible than rumor R_2 according to publicly available evidence, and both rumor proportions are at "reasonable" levels, then we should have $\frac{n_{R_1}}{N_{R_1}} > \frac{n_{R_2}}{N_{R_2}}$.

These design criteria are meant to lead us towards defining what it means for a rumor to be in proportion to its evidence. The criteria state some of the properties that we expect of such "reasonable" rumor proportions, both in terms of the proportions' dynamics and the proportions' relative values across different rumors.

Quantifying Evidence

To make these criteria actionable, a critical component must be specified in more detail. How do we know what constitutes evidence for or against a rumor? How can we measure when one rumor is more plausible than another? To make these determinations, we appeal to the dominant mathematical framework for quantifying evidence and reasoning about uncertainty, Bayesian reasoning. Bayesian reasoning is both a practical toolkit deployed by statisticians to quantify their uncertainty about world and the foremost computational theory of human belief formation in computational cognitive science [20, 21].

The technique is to relate propositions about the world to available information through statistical models of the structure of the situation. Bayesian reasoning represents beliefs about the world as numeric values between 0 and 1, and manipulates those beliefs according to Bayes's Rule. For example, we might have the following model of the Seattle Housing Crisis. If "there is a housing bubble in Seattle that will soon pop" then with high probability we should see rental prices stabilizing or starting to decrease. Let the symbol E_1 refer to this possible consequence. The postulated probabilistic relationship between these two propositions can then be expressed $P(E_1 \mid R) = 0.9$, meaning the probability of E given R is the case is 0.9. 0.9 in this case is a subjective probability judgment that is part of our assumed statistical model. We might also have other consequences to incorporate into our model. Perhaps if there is not an imminent housing market crash in Seattle, then we would expect real estate companies to be creating new development projects, which we could call E_2, and formalize with another conditional probability: $P(E_2 \mid \neg R)$, where $\neg R$ is the logical negation of R. Bayes's Rule then provides us a way to compute a probability judgment on the proposition R given we have observed the two pieces of evidence E_1 and E_2: $P(R \mid E_1, E_2) = \frac{P(E_1, E_2 \mid R)P(R)}{P(E_1, E_2)}$. This probability judgment is called the *posterior probability* of R given E_1 and E_2, with $P(E_1, E_2 \mid R)$ called the *likelihood* of E_1 and E_2 given R; $P(R)$ the *prior probability* of R; and $P(E_1, E_2)$ the *marginal likelihood* of E_1 and E_2. We denote $E = \{E_1, E_2, \ldots\}$ to represent the set of publicly available evidence related to the rumor R. $P(R \mid E)$ is then a posterior probability that the rumor R is true given the available public evidence E.

Mathematical Formulation

Given these mathematical tools for reasoning about uncertainty, the criteria we gave above relatively informally can be further formalized as a mathematical control problem. We can view a system designed to manage rumors as a function $f : [0, 1] \rightarrow [0, 1]$ that takes the public evidence probability $P(R|E)$ as input, and yields a proportion of users ($p_R = \frac{n_R}{N}$) who believe the rumor R as output. The criteria we described in the last section can be formally represented within this general framework as placing a constraint on f to be a strictly monotonically increasing function, defined everywhere on the input space. Alternative goals can also be specified in this framework, but violate these constraints on f. Preventing just false rumors corresponds to attempting to design a system in which $f(0) = 0$, and f is undefined on all other inputs. Preventing all rumors corresponds to a function $f(p) = 0$ for all $p < 1$, which is not strictly monotonic, and is also undefined when $p = 1$. Determining f in a particular application can depend on the cost of false rumors or on the benefit of true rumors. In the following we will explore a particular case that assumes a balance between these costs and benefits.

5 RUMORS IN PROPORTION TO EVIDENCE

Could a system that satisfies the design criteria in the last section ever be achievable? One simple function that satisfies the constraints given is the identity function, $f(P(R \mid E)) = P(R \mid E)$, i.e., aiming to ensure rumor proportions are equal to their level of plausibility, $\frac{n_R}{N_R} = P(R \mid E)$. This particular function also has additional beneficial properties. There are simple behavioral mechanisms of rumor propagation that maintain this rumor proportion as an invariant property over time. In other words, adding superscripts to indicate change over time, there are simple mechanisms that maintain under certain assumptions: If $p_R^{(t)} = P(R \mid E)$, then $E\left[p_R^{(t+1)}\right] = P(R \mid E)$. The same invariant includes a property of adaptive proportions that respond to new evidence: If $p_R^{(t)} = P(R \mid E^{(t)})$, then $E\left[p_R^{(t+1)}\right] = P(R \mid E^{(t+1)})$.

A simple "social sampling" behavior provides a motivating example of one such mechanism [11, 26]. The social sampling procedure consists of people who are undecided about their beliefs sampling proposed beliefs uniformly at random from people who are decided, and then assessing whether to accept the received proposed belief according to evidence. In certain environments, this simple procedure of people learning from each other provably yields the condition $E[p_R] = P(R \mid E)$ [26]. The reason that this procedure maintains this property as invariant once it has been achieved, and while no new evidence is presented, is easy to see. Suppose we have that $p_R = P(R \mid E)$ among people with decided beliefs. When a group of new people enters the conversation and samples among those with decided beliefs, the probability of considering proposition R as true under uniformly random sampling will be exactly equal to the proportion of people who believe it, i.e. $p_R = P(R \mid E)$. As long as the acceptance mechanism for beliefs under consideration does not introduce bias, the proportion of people who accept belief in R will then also be p_R. The property of adapting to new evidence obtains for similar reasons.

Several pieces of evidence support the plausibility of some such social sampling mechanism as an approximate description of human behavior, and support the feasibility of keeping rumors in proportion to $P(R \mid E)$. The evidence come from three places. First, many studies in cognitive science have shown that people display Bayesian updating on average in aggregate [20, 21, 45], and recent work has argued that these results may be best explained by sample-based cognitive mechanisms [47]. Second, some work has directly tested social sampling models as models of behavior outside of the rumor context [26, 28]. Third, several papers in the literature on rumors demonstrate discrete belief propagation and also support the feasibility of both self-correcting dynamics and maintaining in-proportion rumors as an invariant. A common qualitative coding of rumors involves an ontology that generally includes four categories: affirmations, denials, comments/neutral statements, and questions/expressions of uncertainty [30, 32, 34, 43, 50, 53]. The fact that affirmations and denials are separated in this ontology from expressions of uncertainty suggests that people maintain three states: undecided, decided yes, or decided no. Other work has suggested that rumor proportions vary over time more than how people express uncertainty about a rumor [54]. A recent result about the structure of the evolution of rumor content is also consistent with sample-based rumor propagation. Researchers showed that prominent threads of rumors tend to change in discrete mutations more by merging different rumor threads [27]. Merging threads involves entertaining and combining multiple initially inconsistent beliefs about the world, and therefore is not possible if we only think of the world as being one way or another at any particular point in time. Finally, several pieces of work in the literature on rumor detection and classification have noted that detection and classification are possible based on features of discussion patterns, such as the presence of questioning or the proportion of affirmations versus denials [32, 41, 42], which indicates a relationship in existing systems between rumor credibility and rumor proportions. Other work has directly documented self-correcting dynamics [4, 54].

This body of evidence shows that even if people do not always manage to keep rumors in proportion to evidence when completely left to their own devices, mechanisms of human cognition and behavior in rumor contexts are likely at least capable of being oriented towards such an end. This observation brings us back to our prior discussion of the interaction between human behavior and structural considerations in achieving design goals within systems for managing rumors. Equipped with a theoretical prototype of behavior, we can interrogate what structural considerations are sufficient for achieving our design goals given that behavior. If we take the social sampling model of behavior as a starting point, then one key structural condition supports keeping rumors in proportion. The condition is that when one person samples a proposed belief from another person, that sample should be drawn uniformly at random from across all people. Uniform sampling is disrupted when any single person has an undue degree of influence in the sense that that person's belief is more likely to be sampled than someone else's. Having equal contributions to discussion from different people therefore promotes collective sensemaking under this mechanism, while having influencers like network hubs or otherwise overly loud people undermines it. In the following section, we make these considerations more concrete in the form of implications for future research and design, and explore other implications of our general approach, setting aside particular behavioral mechanisms.

6 IMPLICATIONS FOR RESEARCH & DESIGN

Our work has several implications for future research and design. The primary outcome is an emphasis on rumor proportion as a "dependent variable", and an aim to quantify and achieve reasonable rumor proportions. Our work suggests research questions such as: How often do we observe reasonable rumor proportions? What factors lead to this property obtaining or failing to obtain?

Measuring Success

In the course of the development of our approach, we have touched upon both direct and indirect ways to measure whether rumors have grown out of proportion, or are in proportion to evidence. Each of these direct and indirect metrics applies to just a single rumor or a pair of rumors. Analyzing a single rumor can be treated as a single sample of

all the rumors in the ecosystem, and each positive conclusion in a single analysis should increase our assessment of the health of the ecosystem.

Direct Measurements. A direct approach is to take a particular rumor as a sample; have experts (e.g., [31]), or perhaps a prediction market [33], make a quantitative judgment about what is a rational probability $P(R \mid E)$ that the rumor is true given public evidence; and to assess the gap between that probability and the proportion of people who express belief in the rumor either through a survey or through observational data analysis of social media content. The disadvantage of the direct approach is the need to specify a precise quantity $P(R \mid E)$, which can be subjective and variable.

Indirect Measurements. There are several indirect approaches that avoid the requirement to specify a precise quantification of intrinsic uncertainty. Each of the design criteria we specified in Section 4 provide a route to measurement. Suppose an analyst has measured rumor proportions themselves using surveys or social media data analysis. One indirect approach is then to assess whether the release of new public evidence tends to increase or decrease rumor proportions. This analysis would consist of looking at a time series of rumor proportions, marking on that time series when new pieces of public evidence such as news articles are released, annotating those pieces of evidence as supporting the rumor or not, and examining how the time series of rumor proportions tends to vary as a function of those events. Similar methodologies are deployed in "event studies" of the dynamics of financial markets.

Another indirect option when no new information is arriving is to examine pairs of rumors. Rumor proportions can be compared and judged according to whether the difference in proportions across a pair of rumors reflects the difference in credibility across that pair. We offer that in a healthy information ecosystem, differences between rumor proportions will tend to reflect differences in credibility, and when this condition does not obtain, there is reason for concern. Concretely, this analysis would consist of eliciting judgments on which of the rumors is more credible according to public evidence, and then checking that the rumor with greater proportions is also the rumor with higher credibility. While this technique does require assessments of evidence, it does not require precisely quantifying $P(R \mid E)$ since only a rank ordering is needed.

A final example of an indirect option is to examine the proportions of one rumor across different contexts. In healthy information ecosystems, the proportion of a rumor on one platform or in one location or within one demographic should be similar to the proportion of that rumor in other platforms, locations, and demographics. Conducting this analysis would take the form of a subgroup analysis of rumor proportions using survey or social media data.

Social Media Design

We now explore how these considerations could impact specific research and design questions in the context of social media. To be concrete, we will frame these implications in terms of a hypothetical Twitter-like social media system. Since our approach is derived from theory, we frame all of our implications in terms of questions that could be investigated empirically. To that end, in the hypothetical social media system we consider, we could ask:

- Given an expert's judgment of the probability that a rumor is true, does the proportion of people who are affirming that rumor on our social media system match that expert probability judgment?
- How often and for what types of rumors are our indirect measurements of healthy proportions satisfied?
- Across all rumors for which a proportion $\frac{n_R}{N}$ of affirmations is observed, do we in retrospect see that a proportion $\frac{n_R}{N}$ of those rumors was true?
- Which features of the social media platform lead rumors to tend to stay in proportion versus to grow out of proportion?
- When some people post much more than others, there can be differences between the total proportion of rumor affirmations as compared to the total proportion of *people* who affirm the rumor. The social sampling model that we gave as an example of a mechanism under which rumors will be kept in proportion predicts that ensuring this gap is small will help keep rumors in proportion. How far does the proportion of affirmations of a rumor diverge from the proportion of people who affirm it in our system of interest? What design choices influence this gap? Does reducing this gap indeed help keep rumors in proportion as predicted by the social sampling model?
- The same social sampling model also suggests ensuring the probability of exposure to content corresponds to the number of people sharing that content. Is the probability of each user being exposed to a rumor affirmation or denial equal to the proportion of people who affirm or deny that rumor in our system of interest? How can this property be achieved, e.g. in search functionality on the system?

Our theoretical investigation also derived specific interventions on the structure of a social media environment aimed at promoting reasonable rumor proportions. The success of these interventions could be measured by observing how the answers to any of the above questions vary as a function of these interventions.

- Does the repost mechanism promote reasonable rumor proportions or disrupt this goal?
- What effect on rumor proportions does showing featured posts from people you do not follow have?
- Which achieves a healthier rumor proportions: a curated feed or a feed of all content?
- What are the effects on rumor proportions of other mechanisms, such as enabling users to disable replies?

Our work also suggests a shift in how to analyze rumor detection and classification methods.

- Similar to focusing on credibility (e.g., [10]), predict the level of evidence supporting a rumor rather than classifying its veracity.
- Focus on calibration of probabilities as a metric rather than accuracy or false positive/negative rates. If we can correctly predict the probability that a rumor is true, then we can assess whether the rumor proportions are reasonable without resorting to expert judgment.

7 EXAMPLES

We now consider examples of two of our proposed analyses. We perform small-scale analyses that are meant to be illustrative but could easily be expanded. These cases demonstrate how our concepts and metrics can have utility independent of the details of their theoretical foundations. We have made the coded data for these example available online.[2]

Climate Change Beliefs

Our first case builds on the example of climate change. Given the scientific consensus on this issue, climate change denial is more productively viewed as a form of misinformation or disinformation than of rumor, but the case serves to illustrate the concepts and methods of our approach. For this case, we collected data by searching for "global warming" on Twitter and analyzing the top ten English language "Latest Tweets" we observe. The two authors of the present paper independently coded each tweet as either indicating belief or disbelief in anthropogenic climate change, or NA (uncertain or other). We merged the independent codes by coming to consensus where disagreements were present. We then compare the proportion of affirmations versus denials in this sample against a benchmark of the aggregate expert opinion, currently estimated to be between 90-100% of climate scientists [2, 13]. Our codes indicate that only 25% of the sampled tweets affirm the fact of climate change, and therefore this discussion is far out of proportion relative to the expert benchmark. It is well-known that aggregate public beliefs about climate change diverge from expert beliefs (e.g., [23]). Our use of this case study is meant to show how this type of analysis could be conducted in the context of a social media platform even if this deviation was previously undiscovered.

[2]https://github.com/pkrafft/Keeping-Rumors-in-Proportion

Seattle Housing Crisis

Our second case builds on our Seattle Housing Crisis example. We collected data from YouTube and the Seattle Times website. For YouTube, we searched for "seattle housing crisis" and analyzed the first video in the search results with more than 100 comments.[3] For the Seattle Times, we look at the comments section of a relevant article.[4] The two authors of this paper independently coded the top ten comments from each source as either indicating whether Seattle is or is not in the midst of a housing bubble that will soon pop, or NA (uncertain or other). We merged the independent codes by coming to consensus where there were disagreements. We then compare the proportion of affirmations versus denials in the sample from YouTube to the sample from Seattle Times. On YouTube we find that 100% of the non-NA-coded sampled comments affirm the rumor. On the Seattle Times, the proportion was 83%. The gap between the rumor proportions in these two samples indicates that at least one of the two systems could be structured in a way that promotes "out of proportion" discussions of rumors. At the same time, the relatively high proportion across both contexts suggests that public evidence might also favor the truth of the rumor.

8 DISCUSSION

Most mathematical treatments of rumor or other forms of collective behavior offer a specification of a particular behavioral mechanism, e.g. in the form of an algorithm or a differential equation. These "descriptive" accounts aim to capture key features of social phenomena in order to explain or predict human behavior using mathematics and data. Our approach is parallel to and distinct from this modeling tradition. Our primary contribution is to offer a framework for reasoning about rumor proportions and their "health". **The key concepts we introduce to this end are (1) a mathematical *measurement device* (Bayesian rumor probabilities) and (2) a *mathematical criterion* (rumor proportions in relation to Bayesian posterior distributions).** To understand the type of contribution we are trying to make, consider an analogy to the study of urban dynamics. Our contribution is akin to the development of a concept such as measuring the degree of segregation in a city. Regardless of what mechanisms underlie the phenomenon, the new metric provides a way to assess a kind of population health. **The measurement device and the mathematical criterion**

[3]"Seattle Housing Bubble - Unusual Surge in Homes Inventory" https://www.youtube.com/watch?v=qtp-ZdWxo8o

[4]"Seattle-area rents drop significantly for first time this decade as new apartments sit empty" *Seattle Times* January 12, 2018

we introduce can be applied regardless of the details of what rumor mechanisms are at play.

Empirical Validation

A reader may worry about the lack of empirical validation of our framework. Checking fit to data and including key explanatory variables are central to accounts that focus on specific mechanistic models. However, our contribution is a framework not a particular model. Fit to specific datasets or veracity of particular mechanistic descriptions aligned with the framework are inadequate evaluation criteria for such theoretical developments. Frameworks are weighed based on their generative, constructive potential—what thoughts and practices they inspire. We demonstrate this potential in our paper with the implications for research and design we enumerate. We do also present a more classic type of "model"—one behavioral mechanism, "social sampling"—in order to lend credibility to the value of our theoretical maneuvers. We presented evidence for the plausibility of this mechanism through qualitative relationships between properties of social sampling and results in existing published studies. We choose a simple mechanism for clarity of illustration. Future work could test through laboratory or field experiments whether our design implications yield productive threads of research.

Utterances versus Individuals

We have already briefly alluded to the difficulties posed by the difference between the number of people who express belief in a rumor versus the number of actual utterances affirming a rumor in a particular environment. This distinction is important in practical analysis of rumors on social media [5], and for the most part we have elided it. Analysts should be careful to mind this gap, and future theoretical or methodological work on estimating proportions of people who believe a rumor could consider how to adjust for sampling bias induced by differences in numbers of expressions.

Model Extensions

Extensions of Social Sampling Mechanism. Although the specific behavioral mechanism of social sampling is not central to our main contributions, extending this model illustrates how the measurements and diagnostics we introduce can be used in richer contexts. Two important factors that are omitted from the social sampling model are interpersonal trust and network structure. The mathematics of these extensions is a straightforward application of similar extensions of existing models of group belief dynamics [16]. Making these extensions changes the dynamics of the social sampling model so that aggregation to Bayesian rumor proportions is no longer guaranteed. These deviations suggest hypotheses that could be tested about what factors could cause rumors to grow out of proportion. One use of these model extensions within our framework is therefore to examine how changes in rumor mechanisms promote or inhibit healthy rumor proportions.

Extensions to Measurements and Diagnostics. Extending the measurements and diagnostics we introduce is somewhat more subtle. How should we conceptualize the notion of "healthy rumor proportions" in light of considerations such as incentive, affect, or trust? Should we judge dangerous rumors more harshly? What about disturbing or offensive rumors? And what of rumors in contexts when scientific knowledge or other expert judgments are suspect? In the present work we avoided these issues by proclaiming a focus on non-political "rumors of fact", but extensions to other situations are desirable. For incentives, existing decision-theoretic models provide a guide for how to modify distributions of beliefs in high-cost or high-gain situations [18]. For disturbing or offensive rumors, we would have to incorporate constraints of the sort discussed in debates about platform moderation (e.g., [19]). To incorporate considerations of public mistrust, we could adjust the sources of who generates a normative baseline probability judgment. Rather than relying on expert judgment we could, for example, define the proper posterior probability to be the one that a lay person presented with a body of evidence would give.

9 CONCLUSIONS

The primary contribution of our work is a *framework for reasoning* that we make actionable with *new tools for measurement and diagnosis.* Although this framework involves a modeling approach, our focus is not on a description of a single "model" as in a specific proposed mechanism of rumor. We do not aim to present the definite end-all-be-all model for all rumors, but rather to offer a point of inspiration and a lens for analysis. The measurement devices and the mathematical criteria we introduce can be applied regardless of the details of what rumor mechanisms are at play. The ultimate recommendation we make for the design of systems to deal with rumors is to attend to the proportion of affirmations of a rumor, and attempt to keep rumor affirmations in proportion to the evidence for the rumor. We offered both direct and indirect ways to diagnose whether this goal is achieved.

10 ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants 1420255 and 1715078, by the Washington Research Foundation, and by a Data Science Environments project award from the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #3835) to the University of Washington eScience

Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] G.W. Allport and L. Postman. 1947. *The Psychology of Rumor.* Henry Holt.

[2] William RL Anderegg, James W Prall, Jacob Harold, and Stephen H Schneider. 2010. Expert credibility in climate change. *Proceedings of the National Academy of Sciences* 107, 27 (2010), 12107–12109.

[3] Susan Anthony. 1973. Anxiety and rumor. *The Journal of Social Psychology* 89, 1 (1973), 91–98.

[4] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 155–168.

[5] Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S Spiro. 2016. How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 466–477.

[6] Elizabeth Bonawitz, Stephanie Denison, Alison Gopnik, and Thomas L Griffiths. 2014. Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology* 74 (2014), 35–65.

[7] Prashant Bordia and Nicholas DiFonzo. 2004. Problem solving in social interactions on the Internet: Rumor as social cognition. *Social Psychology Quarterly* 67, 1 (2004), 33–49.

[8] T. Caplow. 1947. Rumors in war. *Social Forces* 25, 3 (1947), 298–302.

[9] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 675–684.

[10] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588.

[11] L Elisa Celis, Peter M Krafft, and Nisheeth K Vishnoi. 2017. A distributed learning dynamics in social groups. In *Proceedings of the ACM Symposium on Principles of Distributed Computing*. ACM, 441–450.

[12] David Coast and Jo Fox. 2015. Rumour and politics. *History Compass* 13, 5 (2015), 222–234.

[13] John Cook, Dana Nuccitelli, Sarah A Green, Mark Richardson, Bärbel Winkler, Rob Painting, Robert Way, Peter Jacobs, and Andrew Skuce. 2013. Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters* 8, 2 (2013), 024024.

[14] Nicholas DiFonzo and Prashant Bordia. 2000. How top PR professionals handle hearsay: Corporate rumors, their effects, and strategies to manage them. *Public Relations Review* 26, 2 (2000), 173–190.

[15] Nicholas DiFonzo, Prashant Bordia, and Ralph L Rosnow. 1994. Reining in rumors. *Organizational Dynamics* 23, 1 (1994), 47–62.

[16] Noah E Friedkin and Eugene C Johnsen. 2011. *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. Cambridge University Press.

[17] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *International Conference on Weblogs and Social Media (ICWSM)*.

[18] John Geweke. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society* (1989), 1317–1339.

[19] Tarleton Gillespie. 2018. Platforms are not intermediaries. *Georgetown Law Technology Review* 2 (2018), 198.

[20] Thomas Griffiths, Charles Kemp, and Joshua Tenenbaum. 2008. Bayesian models of cognition. *Cambridge Handbook of Computational Cognitive Modeling* (2008).

[21] Thomas L Griffiths and Joshua B Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological Science* 17, 9 (2006), 767–773.

[22] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get back! You don't know me like that: The social mediation of fact checking interventions in Twitter conversations. In *Proceedings of the International Conference on Weblogs and Social Media*.

[23] Peter D Howe, Matto Mildenberger, Jennifer R Marlon, and Anthony Leiserowitz. 2015. Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change* 5, 6 (2015), 596.

[24] Caroline Jack. 2017. *Lexicon of lies: Terms for problematic information*. Technical Report. Data & Society Research Institute.

[25] Michael L Kalish, Thomas L Griffiths, and Stephan Lewandowsky. 2007. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review* 14, 2 (2007), 288–294.

[26] Peter Krafft. 2017. *A Rational Choice Framework for Collective Behavior*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[27] Peter Krafft, Kaitlyn Zhou, Isabelle Edwards, Kate Starbird, and Emma S Spiro. 2017. Centralized, parallel, and distributed information processing during collective sensemaking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2976–2987.

[28] Coco Krumme, Manuel Cebrian, Galen Pickard, and Sandy Pentland. 2012. Quantifying social influence in an online cultural market. *PLoS One* 7, 5 (2012), e33785.

[29] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLoS one* 12, 1 (2017), e0168344.

[30] Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Using Gaussian processes for rumour stance classification in social media. *arXiv preprint arXiv:1609.01962* (2016).

[31] Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25, 5 (2014), 1106–1115.

[32] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we RT?. In *Proceedings of the First Workshop on Social Media Analytics (SOMA)*.

[33] David M Pennock, Steve Lawrence, C Lee Giles, and Finn Årup Nielsen. 2001. The real power of artificial markets. *Science* 291, 5506 (2001), 987–988.

[34] Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology* 16, 3 (2013), 197–214.

[35] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[36] Ralph L Rosnow. 1980. Psychology of rumor reconsidered. *Psychological Bulletin* 87, 3 (1980).

[37] Ralph L Rosnow. 1988. Rumor as communication: A contextualist approach. *Journal of Communication* 38, 1 (1988), 12–28.

[38] Ralph L Rosnow. 1991. Inside rumor: A personal journey. *American Psychologist* 46, 5 (1991), 484.

[39] Adam Sanborn and Thomas L Griffiths. 2008. Markov Chain Monte Carlo with people. In *Advances in Neural Information Processing Systems*. 1265–1272.

[40] Tamotsu Shibutani. 1966. *Improvised News: A Sociological Study of Rumor*. The Bobbs-Merrill Company, Inc., Indianapolis, New York.

[41] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society* 19, 8 (2017), 1214–1235.

[42] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings* (2014).

[43] Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. Could this be true?: I think so! Expressed uncertainty in online rumoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 360–371.

[44] Cass R Sunstein. 2014. *On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can Be Done*. Princeton University Press.

[45] Joshua Brett Tenenbaum. 1999. *A Bayesian Framework for Concept Learning*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[46] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. 2010. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 1817–1820.

[47] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science* 38, 4 (2014), 599–637.

[48] Edward Vul and Harold Pashler. 2008. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science* 19, 7 (2008), 645–647.

[49] Helena Webb, Pete Burnap, Rob Procter, Omer Rana, Bernd Carsten Stahl, Matthew Williams, William Housley, Adam Edwards, and Marina Jirotka. 2016. Digital wildfires: Propagation, verification, regulation, and responsible innovation. *ACM Transactions on Information Systems (TOIS)* 34, 3 (2016), 15.

[50] Li Zeng, Kate Starbird, and Emma S Spiro. 2016. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Tenth International AAAI Conference on Web and Social Media*.

[51] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 603–612.

[52] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.

[53] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 32.

[54] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One* 11, 3 (2016), e0150989.