Internet Outages, the Eyewitness Accounts: Analysis of the Outages Mailing List

Ritwik Banerjee[†] Abbas Razaghpanah[†] Luis Chiang[†] Akassh Mishra[†]
Vyas Sekar[°] Yejin Choi[‡] Phillipa Gill[†]

[†]Stony Brook University, [°]Carnegie Mellon University, [‡]University of Washington

Abstract. Understanding network reliability and outages is critical to the "health" of the Internet infrastructure. Unfortunately, our ability to analyze Internet outages has been hampered by the lack of access to public information from key players. In this paper, we leverage a somewhat unconventional dataset to analyze Internet reliability—the outages mailing list. The mailing list is an avenue for network operators to share information and insights about widespread outages. Using this unique dataset, we perform a first-of-its-kind longitudinal analysis of Internet outages from 2006 to 2013 using text mining and natural language processing techniques. We observe several interesting aspects of Internet outages: a large number of application and mobility issues that impact users, a rise in content, mobile issues, and discussion of large-scale DDoS attacks in recent years.

1 Introduction

As an increasing number of critical services rely on the Internet, network outages can cause significant societal and economic impact [10, 18]. Indeed, this importance can be seen when network failures such as cloud computing outages [9], BGP interceptions [14], and large scale DDoS attacks (e.g., [1, 3]) make headlines in the popular press. By some estimates, data center network outages can lead to losses of more than $500,000 per incident on average [34], while costs of WAN failures are more challenging to quantify [8]. Thus, there are a large number of past and ongoing efforts to detect and mitigate network outages, including work on novel root cause analysis techniques [24, 27], and better network debugging tools [5, 11, 20, 30, 41].

While there are several efforts, as mentioned above, to minimize the impact of network outages, there is unfortunately a critical dearth of studies that systematically *understand* network outages. In part, our understanding of outages and network reliability is hampered by the reluctance on the part of network operators to release data due to policy requirements; e.g., even though the FCC maintains a network outage reports system and mandates that network operators provide true estimates, the data is confidential given its sensitive nature [2]. Furthermore, providers have natural economic concerns that such studies may reflect poorly on them and thus impact revenues. As such, the few studies that obtain data from networks are only able to offer insights from a single vantage point such as an academic WAN [43], data center [22] or backbone ISP [32].

Our work is an attempt to bridge this critical gap in our understanding of network reliability. For instance, we would like to understand if specific Internet service providers (e.g., access vs. tier-1), protocols (e.g., DNS vs. BGP), network locations (e.g., specific PoPs or co-location points), or content providers (e.g., web hosting services) are more

likely to be involved in network outages. Such an understanding can help network operators and architects focus their resources on making Internet services more robust. For example, providers who know that specific hosting services or protocols are prone to outages can proactively work around these known hotspots.

Toward this goal, we leverage an underutilized dataset: the *outages mailing list* [38] to answer the above types questions. The mailing list serves as a venue for operators to announce and debug network failures. The outages list tends to have some bias towards North American network operators self-reporting outages perceived as 'high impact'. Despite this bias, the dataset also has attributes that are lacking, or only met in isolation in other data sets which can help illuminate different facets of network failures:

Semantic context. Posts contain rich semantic information about what happened during the outage, in contrast to technical data which often requires starting from low-level measurements and inferring whether an event incurred real-world impact.

Interdomain coverage. The mailing list provides an overview of network failures that transcend network boundaries rather than focusing on the point-of-view and failures experienced by a single network.

Longitudinal view. The outages list has been maintained since 2006 offering an unprecedented view of Internet reliability issues discussed by operators over time.

The rich semantic and natural language information contained in the list also presents a challenge in terms of analyzing the outages mailing list. To address this challenge, we turn to natural language processing (NLP), text mining, and machine learning (ML) techniques in order to automatically categorize the posts and threads in the mailing list. However, naively applying these techniques "out of the box" does a poor job of identifying useful semantic information (e.g., Level 3 would naively be considered two words). Thus, we use a careful synthesis of domain knowledge and NLP/ML techniques to extract meaningful keywords to build a classification algorithm to categorize content along two dimensions: (1) type of outage (e.g., attack vs. congestion vs. fiber cut) and (2) the type of entity involved (e.g., cloud provider vs. ISP).

Our analysis reveals the following insights:

User issues dominate. The list is dominated by issues with user-facing components such as misconfigurations and issues with application servers and mobile networks. In terms of entities, networks providing service to users such as access and mobile networks are also prevalent.

Content and mobile issues are on the rise. Starting in 2009, we see a large fraction of threads related to application server problems and content provider networks. These issues tend to relate to common service providers such as Google, Facebook, Netflix. Mobile-centric issues have also increased by 15% over the past 7 years.

Attacks and censorship are relatively rare. There is less discussion of security issues and censorship in the dataset. However, notable incidents like censorship in Syria and large DNS-amplification-based DDoS attacks (*e.g.,* [35]) did get the attention of the community with a significant increase in posts containing the keyword DNS spiking in 2012-2013.[1]

[1] DNS was used to amplify botnet attacks over this period.

Contributions and Roadmap: This paper makes the following contributions: (1) Performing an initial analysis of the outages mailing list to understand Internet outages (§2); (2) A careful application of text mining, NLP, and machine learning techniques to extract useful semantic information from this dataset (§3,§4); (3) Shedding light on the types of outages and the key entities involved in these outages over time (§5). Finally, we discuss related work in §6 and conclude in §7.

First Email	Sep 29, 2006
Last Email (in dataset)	Dec 31, 2013
Number of Posts	6,566
Number of Threads	2,054
Number of Replies	4,163
Number of Contributors	1,194

Table 1: Summary of the Outages Mailing List Dataset

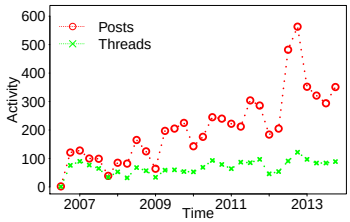


Fig. 1: Outages Mailing List Activity per Quarter.

2 Dataset

In this section, we provide background about the mailing list and our dataset (§2.1), and limitations of using the mailing list to analyze network failures (§2.2).

2.1 About the Outages Mailing List

The outages mailing list reports outages related to failures of major communications infrastructure components. It intends to share information so that network operators and end users can assess and respond to major outages. The list contains outage reports as well as post-mortem analysis and discussions on troubleshooting.

We analyze a snapshot of the outages mailing list taken on December 31, 2013 containing threads since its inception in 2006. Our dataset is summarized in Table 1. It contains over seven years of discussion on the mailing list. This discussion is organized into 2,054 threads, with a total of 6,566 individual posts. Note that the number of posts is higher than the number of threads and replies combined since it also includes emails that are not part of a thread (e.g. ”unsubscribe” emails). A total of 1,194 individuals (identified by e-mail addresses) contributed to the discussions.

Activity on the mailing list shows an upwards trend since it was started in 2006. Figure 1 shows quarterly activity on the list in terms of the number of threads and posts. The amount of activity on the list shows a periodic trends with less activity in Q4 which includes the holiday season. We also observe a spike in posts towards the end of 2012 which can be attributed to discussions arising from Hurricane Sandy.

2.2 Limitations

While the mailing list provides a unique view of failures which had observable impact over the past seven years, it also has some limitations. The data is biased towards North American operators and Internet providers since many of the users are US-based system administrators and the forum itself is hosted in North America. Moreover, we are biased towards incidents which transcend network boundaries as incidents which remain internal to a network are unlikely to be posted. Further, the list does not contain technical

information about the underlying root cause, and indeed some posts lack a clear root cause. Finally, while the list contains failures that impacted users, there is some selection bias in terms of failures that users report to the list (*e.g.,* the aforementioned North American bias, and bias towards networks upstream of networks whose operators are more active in the list). Despite these limitations, the data contained in the mailing list is valuable because it presents a longitudinal and cross-provider view of failures that had real world impact on the Internet.

3 Keyword Analysis

In this section, we discuss how we extract keywords from the e-mail postings (§3.1) and present preliminary analysis of topics over time (§3.2).

3.1 Data Preprocessing

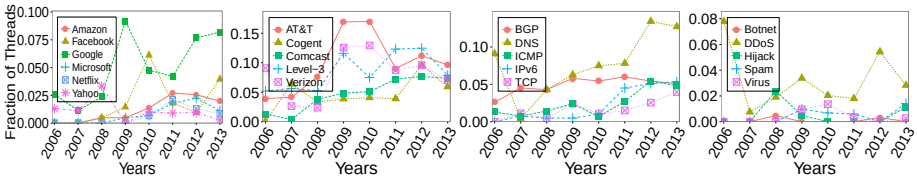
The fact that e-mail postings are comprised of natural language text means that they are rich with semantic information underlying the failure, but also presents a challenge in terms of automatically parsing and processing the data. To address this challenge we employ techniques from text mining and natural language processing (NLP).

Step 1: Collate threads. In general, we consider the dataset at the level of *threads*. Each thread consists of the set of e-mail messages (posts) in the thread. For each thread we extract relevant terms and phrases after removing quoted text (text from previous emails in the thread included in each email) from its posts.

Step 2: Remove spurious data and stop-words. We first discard spurious data contained in the posts. This included identifying e-mail signatures used by posters which contributed to terms and phrases unrelated to the content of the thread. We also extract traceroute measurements which are often contained in posts at this point. While traceroutes are useful for debugging, it is difficult to identify the root cause of an incident via automated analysis of the traceroutes, since the list contains posts on a variety of topics. Thus, we focus on the natural language content of the messages in this paper. We leverage a list of 572 stop words (*e.g.,* articles, prepositions and pronouns) obtained from the SMART information retrieval system [37]. Punctuations are also removed.

The remaining words are lemmatized (the process of grouping together the different inflected forms of a word) using the Stanford CoreNLP toolkit [4] so they can be analyzed as a single item. For example, determining that "walk", "walked" and "walking" are all forms of the same verb: "to walk". Note that the simple stemming (i.e., walking $\rightarrow$ walk) does not suffice as it cannot differentiate the parts of speech based on context: *e.g.,* when the term "meeting" acts as a verb: "we are meeting tomorrow" *vs.* a noun "let's go to the meeting". Lemmatization, on the other hand, can identify these contextual differences. Additionally, we filter out words with term-frequency inverse document frequency (*tf-idf*) values less than 0.122. Low *tf-idf* indicates that the word is very common throughout the dataset [36]. The threshold was chosen such that it filtered out the bottom 25% of terms in terms of tf-idf value.

Step 3: Extract nouns and named entities. To obtain additional information about terms contained in the e-mail messages, we use the Stanford part-of-speech tagger [42] and named-entity recognizer [21]. These tools allow us to identify nouns as well as named entities (*e.g.,* identifying "Los Angeles" as a single entity). This process, how-



(a) Content providers (b) ISPs (c) Protocols (d) Security

Fig. 2: Keyword trends over the years in the outages mailing list.

Root cause of outage
Congestion, Censorship, Fiber Cut, Device Failure, Natural Disaster, Routing, App. Misconfiguration, Mobile Data Network, DNS Resolution, App. Server Down, Attack, Power Outage
Entities involved
ISP, Cloud Provider, Content Delivery Network, Mobile, Email, Access, Content

Table 2: Summary of categories

ever, is incomplete for domain-specific entities found in networking-related e-mails. This problem is particularly acute for organization names (*e.g.,* "Level 3"). Instead of retraining the named entity recognition system – a process that would have required extensive human annotation – we leverage Wikipedia to improve named entity recognition for networking entities. We use the simple heuristic that if a term is a capitalized noun, we search for this term or phrase in Wikipedia. If we identify a page which contains this term as the title, we check that the page is a subcategory of the "Telecommunications companies" category. If the page is in this category, we determine that the term is likely the name of a relevant organization. For multi-word entities such as "Time Warner Cable", we consider noun sequences instead of a single term to search for Wikipedia titles.

3.2 Keyword trends

As a first step, we consider keyword trends to understand failures discussed in the list (Figure 2). We focus on keywords in four categories: content providers, ISPs, protocols, and security. For each category we select 5-6 potentially interesting keywords. Among content providers, Google being the most popular, is more heavily discussed than others. In terms of ISPs, AT&T, Verizon and Level-3 are the most frequently discussed, with an upward trend in ISP-related discussion over time. In terms of protocols, BGP and DNS dominate, with DNS experiencing a sharp uptick in discussions in 2012-2013. Our analysis based on binary classifiers (explained in §4) shows that this is due to a more than twofold increase of DNS-related issues among access (from 3.3% in 2011 to 7.0% in 2012) and content providers (0.9% in 2011 to 2.2% in 2012). Finally, we observe DDoS as the most prevalent term related to security. It comprises nearly 8% of posts in 2006 (note that we only have two months of data in 2006) and surges again to 5.5% in 2012 as a result of large DDoS attacks which occurred that year (*e.g.,* [35]).

4 Classification methodology

The terms and phrases extracted in our initial processing give a high-level view of the discussions on the mailing list. In this section, we discuss a classification methodology to help us systematically categorize the outages over time.

Conceptually, we can categorize a network outage along two orthogonal dimensions: (1) *type of the outage* (e.g., fiber cut), and (2) *entities involved in the outage* (e.g., access ISPs). Table 2 summarizes the specific categories of types and entities of interest.[2] Thus, our goal is to automatically characterize each outage e-mail thread into categories along these dimensions. Next, we describe how we designed such a classifier.

Labeling: As a first step toward automatic classification, we created a simple website to enable us and our collaborators to manually label a small random sample of the posts along the above two dimensions. We had 5 volunteers, each labeling around 30 threads. To validate that our manual annotations were consistent, we use the Fleiss' κ metric [29]; the κ value was 0.75 for entities and 0.5 for the outage types. To put this in perspective, 0.748 is considered very good and 0.48 is considered a "moderate agreement" [29]. Given this confidence, we use these manual labels to bootstrap our learning process described below.

Choice of algorithm: Our initial intuition was to formulate this as a semi-supervised clustering problem [6,17,46]. That is, we use the labeled data to bootstrap the clustering process, learn features of the identified clusters, and then iteratively refine the clusters. However, we found that the *training error* was quite high (i.e., low F-score on the labeled set). The primary reason for this is the well-known *class imbalance* problem — most real-world datasets are skewed with a small number of classes contributing the most "probability mass". The small number of training samples meant this problem was especially serious in our context.

Given this insight, we reformulated the semi-supervised clustering as a *classification* problem. While classification by itself is not immune to class imbalances, it can be made robust using two well-known ideas: (1) learning multiple binary classifiers and (2) suitable resampling [23, 28, 44]. For (1), instead of partitioning the dataset into N categories, we learn a "concept" for each category independently; i.e., a binary classifier trying to determine whether a thread belongs in a particular category or not. For (2), we setup the training with undersampling the majority class and/or oversampling the minority class to make the training data more balanced.

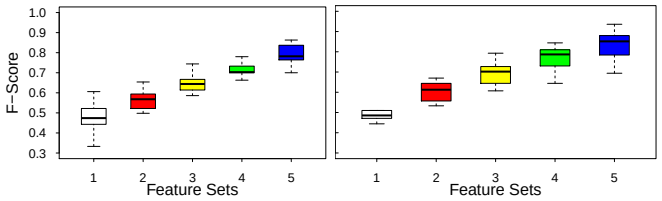
We chose a linear-kernel SVM for classification using the LibLINEAR toolkit [19] which performed well in terms of both accuracy and speed. We evaluate the goodness of the learning step using a standard leave-one-out cross-validation and compute the F-score, which is the harmonic mean of precision and recall values [31]. Next, we describe the features provided to the machine learning algorithm.

Feature selection and refinement: The naïve way to set up a NLP classification is to use a standard "bag-of-words" approach—extract words appearing in the entire dataset and create a binary feature vector for each thread indicating whether a specific keyword appears in it. This approach, however, yields very poor results on two fronts. First, while natural language text contains some terms relevant to the outage, it mostly contains English words which are not relevant to the topic and simple filtering steps such as removing stop words (e.g., "the") do not alleviate this problem. Second, this naïve set of features produces a high-dimensional feature space creating more noise.

[2] We do not claim that this list is exhaustive; it represents a pragmatic set we chose based on a combination of domain knowledge and manually inspecting a sample of the dataset.

Root cause of outage	**Entities involved**
1. Unigrams	1. Unigrams + bigrams (nouns)
2. Unigrams + bigrams	2. Unigrams + bigrams (nouns) + positional weights
3. Nouns	3. Nouns + named entities
4. Unigrams + bigrams (nouns)	4. Named entities
5. Unigrams + bigrams (nouns) + positional weights	5. Named entities + Wikipedia category information

Table 3: Summary of feature sets used to improve the performance of the classifiers

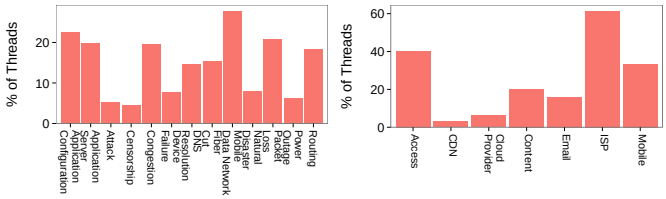


(a) Type of outage (b) Entity

Fig. 3: F-score of classification results using different feature sets. A higher F-score implies better accuracy and the result shows the effect of our iterative feature refinement process. Table 3 summarizes the feature sets.

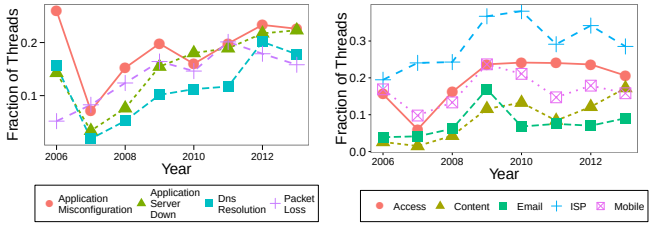
Thus, we had to take further care in selecting the *feature set* using a combination of domain knowledge and manual inspection as described below. First, since most terms associated with our labels are likely to be nouns, we used a part-of-speech tagger [42] to filter out verbs and adjectives. Second, based on manual inspection, we found that terms in the title of the thread, or near the end of the thread were more informative and thus we experimented with weighing these terms higher. The reason is that the issues are mostly resolved towards the end of the discussion and the terms used are more pertinent to the issue. Third, to identify the entities involved, we further prune the features using a named-entity recognition system [21]. While this step retains good features (*i.e.,* words or phrases recognized as entities), it does not provide any semantic information about them. To this end, we used Wikipedia category information to glean such semantic associations. We collected 20,105 Wikipedia pages under the category "Computer Networking", and weighted the features according to whether they occur in pages under relevant subcategories (*e.g.,* "Akamai" under "Content Delivery Network"). We thus designed feature vectors with relevant entities, and weighted them according to their type. (Note that these three steps are in addition to the preprocessing in §3 that was less analysis-specific.)

Table 3 summarizes the different sets of terms we used and Figure 3 shows how the F-score improves as we add better features. The final features selected differ between the type and entity classifiers; i.e., nouns weighed by their position in the thread performing best for *root cause* and a combination of named entities+Wikipedia category information for the *entities involved*. With these features the mean F-score of the classifiers was 78.8% for *root cause* and 82.9% for *entities involved*. For multi-class classification tasks for which human annotation κ scores are in the range of 0.5 – 0.78, these results can be considered as reasonably high. Given the relatively small training data set and the succinct nature of the mailing list posts, the resulting performance is very promising, especially for domains for which a large number of user contributed posts are available for analysis.



(a) Type of outage (b) Entity

Fig. 4: Percentage of threads classified into each class



(a) Types of outages (b) Entities

Fig. 5: Distribution of topics over time for topics that change by at least 10%

.

Finally, one concern with our binary classification approach is the risk that a given thread falls in multiple classes. Fortunately, we found that the majority (>80%) of threads had at most 1 label (not shown).[3]

5 Characterizing the causes of failures

Next, we use the classification methodology from the previous section to analyze common causes and types of outages discussed in the mailing list. Figure 4 shows the fraction of threads classified based on their outage and entity types.

Outage types are dominated by user-observed issues. We find that the majority of threads are placed in categories that indicate user impact. For outage type, mobile data network issues, application server, and application configuration issues dominate, comprising 28%, 20%, and 23% of the data respectively. Upon closer inspection we find common terms in the application clusters related to load balancing, server errors, and browsers (along with common applications like Facebook). For mobile, we found mobile network operators like AT&T, Sprint, and Verizon were common keywords. After issues faced by users, topics tend to be related to more operational issues such as congestion, packet loss, and routing. Issues related to attacks, censorship, natural disasters, and power outages are less common.

Dominant entities are access, ISP and mobile networks. Figure 4 highlights the prevalence of ISPs, access networks and mobile networks as entities involved in the outages. Overall, errors in application-specific entities like CDNs, e-mail, cloud and content providers were less prevalent in the mailing list discussions. Keywords in the

[3] The few threads with multiple labels were often related; e.g, congestion and packet loss or mobile + ISP.

ISP: Level 3	**Content: Facebook**	**Mobile: AT&T**
Class Label (% of threads)	Class Label (% of threads)	Class Label (% of threads)
Congestion (15.1)	App. Server (14.5)	Mobile Data Networks (26.0)
Packet Loss (14.7)	Mobile Data Netw. (12.9)	App. Misconfiguration (12.0)
Routing (14.2)	App. Misconfiguration (12.1)	Packet Loss (9.6)

Table 4: Correlation between entity keywords, and cause of outage

access category tended to include access network providers like Verizon, Comcast, and Time Warner as well as issues like latency, time outs, and fiber cuts.

Content and mobile issues are on the rise. Figure 5 shows the breakdown of topics by year for outage and entity types, respectively. Starting in 2009 we see the emergence of Content providers as an entity that is commonly discussed in the mailing list. That same year we begin to see more posts related to application misconfigurations. We also observe a corresponding increase in issues related to mobile data.

Correlating keywords and associated outage types. We revisit some of the keywords observed in Figure 2 and consider the top outage types for threads containing these keywords in Table 4. We consider keywords related to specific entities in three broad classes: ISP (Level 3), content provider (Facebook), and mobile ISP (AT&T). We find that threads containing Level 3 (and other ISPs we consider), tend to relate to operational issues for the network such as congestion, packet loss and routing incidents. In contrast, Facebook and AT&T tend to be discussed in relation to application server/misconfiguration issues and mobile data network issues. Interestingly, we also observe Facebook in threads related to mobile data network issues, possibly related to mobile users having trouble reaching the site. Similarly, AT&T is mentioned in threads related to application misconfigurations *e.g.,* application specific CDN configurations that may impact users on a specific ISP.

High impact events. Finally, we investigate two incidents which explain spikes in posting in 2012. Among threads with the longest duration and most replies, are those related to a series of large-scale DNS amplification DDoS attacks in September 2012 [1, 3]. Threads related to the issue reported performance problems in DNS servers that, as a result of misconfiguration, were acting as open resolvers. These servers were inadvertently flooding targets with large DNS responses, which in turn degraded performance for legitimate DNS queries [40].

Another spike in activity is related to a widespread outage in late October 2012, experienced by users of Windstream, a large ISP in the United States. Users in multiple areas (mainly in the north and northeastern US) experienced outages due to a fiber-cut caused by Hurricane Sandy [45]. Many outages around that time–related to Hurricane Sandy–also contributed to the increase in mailing list activity during fall 2012 [15]. We manually verified that these high-impact events were correctly classified by the machine learning method in terms of both the type of outage and the entities involved.

6 Related work

Intradomain reliability. Network reliability has been considered in a variety of networks ranging from an academic WAN [43] and ISPs [32, 47] to data centers [22, 39] using a variety of data sources. Some monitor properties of intradomain routing protocol such as OSPF Link State Advertisements (LSAs), which can indicate instability

or unavailability of network links, or IS-IS messages which require specialized infrastructure for monitoring. More recently, there has been interest in using syslog–which is ubiquitous in many networks–to infer and study network failures. Because these studies rely on protocol and logging messages to infer the state of the netwotk, they have a hard time inferring real user impact. Further, in many cases the network is an important part of the business which makes revealing failures unattractive.

Interdomain reliability. A variety of techniques have been employed to understand reliability at the interdomain level, including ongoing probing and monitoring efforts [26] and crowdsourcing measurements from a large population of P2P users [13]. However, characterization of the Internet's reliability at this level has been hindered by the limited view of the system provided by publicly available datasets (*e.g.,* BGP feeds).

Application layer and user-reported reliability. Network level failures do not always imply application layer or user-observed impact. There have been some studies that specifically try to address this using different techniques. Web application reliability was measured by monitoring Web client connections [33] to determine if failures were primarily client or server-related. Netmedic [25] analyzes correlations between application servers that fail in an enterprise network to understand root cause. In the context of cloud computing, Benson *et al.* attempt to mine threads from customer forums of an IaaS cloud provider [7] to identify problems users face when using cloud computing. This work is similar to our own in that it attempts to gather data from naturally arising user discussions, however, their work takes a more focused view considering only failures of a specific cloud provider.

Concurrently to our study, Dimitropoulos and Djatmiko also recognized the potential of mailing lists as a dataset [16]. However, their analysis is orthogonal to ours, which focuses more on how to apply NLP to exploit the semantics of these datasets and understand them at-scale.

7 Conclusions

In this paper, we explore an operator-run mailing list to understand reliability issues spanning multiple networks over a period of 7 years. Our main observations are that the list is primarily used for discussing issues raised by users (*e.g.,* application and mobile data issues) and that content services are on the rise in terms of discussion threads.

The mailing list data presents only one of many natural language resources that can be used to understand network reliability and the methodology applied in this paper will hopefully inspire further analysis of natural language network datasets (*e.g.,* forums [7] and trouble shooting tickets [12]) and mailing lists such as NANOG. Text-based analysis may also be combined with empirical troubleshooting approaches (*e.g.,* Hubble [26], LIFEGUARD [27]) to provide a more complete view of network reliability when directly measured data is scarce, incomplete, or unavailable.

References

1. Deep Inside a DNS Amplification DDoS Attack. `http://blog.cloudflare.com/deep-inside-a-dns-amplification-ddos-attack`.
2. FCC NETWORK OUTAGE REPORTING SYSTEM (NORS). `http://transition.fcc.gov/pshs/services/cip/nors/nors.html`.

3. Spamhaus DDoS grows to Internet-threatening size. http://arstechnica.com/security/2013/03/spamhaus-ddos-grows-to-internet-threatening-size/.
4. Stanford corenlp. http://nlp.stanford.edu/software/corenlp.shtml.
5. R. Alimi, Y. Wang, and Y. R. Yang. Shadow configuration as a network management primitive. In *SIGCOMM*, 2008.
6. S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine Learning (ICML)*, volume 2, pages 27–34, 2002.
7. T. Benson, S. Sahu, A. Akella, and A. Shaikh. A first look at problems in the cloud. In *HotCloud*, 2010.
8. M. Brandenburg. Determining the impact of wide area network outages. http://searchenterprisewan.techtarget.com/feature/Determining-the-impact-of-wide-area-network-outages.
9. J. Brodkin. Amazon ec2 outage calls 'availability zones' into question, 2011. http://www.networkworld.com/news/2011/042111-amazon-ec2-zones.html.
10. Growing business dependence on the internet: New risks require CEO action, 2007. http://businessroundtable.org/sites/default/files/200709_Growing_Business_Dependence_on_the_Internet.pdf.
11. X. Chen, Y. Mao, Z. M. Mao, and K. van de Merwe. Declarative configuration management for complex and dynamic networks. In *CoNEXT*, 2010.
12. Y.-C. Cheng, J. Bellardo, P. Benko, A. Snoeren, G. Voelker, and S. Savage. Jigsaw: Solving the puzzle of enterprise 802.11 analysis. In *SIGCOMM*, 2006.
13. D. Choffnes, F. Bustamante, and Z. Ge. Crowdsourcing service-level network event detection. In *SIGCOMM*, 2010.
14. J. Cowie. Renesys blog: China's 18-minute mystery. http://www.renesys.com/blog/2010/11/chinas-18-minute-mystery.shtml.
15. B. Darrow. Superstorm Sandy wreaks havoc on internet infrastructure, 2012. https://gigaom.com/2012/10/30/superstorm-sandy-wreaks-havoc-on-internet-infrastructure/.
16. X. Dimitropoulos and M. Djatmiko. Analysis of outage posts in the nanog and outages mailing lists, 2013. https://tnc2013.terena.org/core/presentation/146.
17. C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proc. KDD*, 2006.
18. S. Dynes, E. Andrijcic, and M. E. Johnson. Costs to the US economy of information infrastructure failures: Estimates from field studies and economic data. In *WEIS*, 2006.
19. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
20. N. Feamster and H. Balakrishnan. Detecting BGP configuration faults with static analysis. In *Sigcomm*, 2005.
21. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. ACL*.
22. P. Gill, N. Jain, and N. Nagappan. Understanding network failures in data centers: Measurement, analysis, and implications. In *SIGCOMM*, 2011.
23. N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Intl Conf. on Artificial Intelligence*. Citeseer, 2000.
24. U. Javed, I. Cunha, D. R. Choffnes, E. Katz-Bassett, T. Anderson, and A. Krishnamurthy. Poiroot: Investigating the root cause of interdomain path changes. In *SIGCOMM*, 2013.
25. S. Kandula, R. Mahajan, P. Verkaik, S. Agarwal, J. Padhye, and P. Bahl. Detailed diagnosis in enterprise networks. In *SIGCOMM*, 2010.
26. E. Katz-Bassett, H. Madhyastha, J. John, A. Krishnamurthy, D. Wetherall, and T. Anderson. Studying black holes in the internet with hubble. In *NSDI*, 2008.

27. E. Katz-Bassett, C. Scott, D. R. Choffnes, I. Cunha, V. Valancius, N. Feamster, H. V. Madhyastha, T. Anderson, and A. Krishnamurthy. LIFEGUARD: Practical repair of persistent route failures. In *SIGCOMM*, 2012.
28. M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *International Conference on Machine Learning (ICML)*, volume 97, pages 179–186, 1997.
29. J. R. Landis, G. G. Koch, et al. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174, 1977.
30. A. Mahimkar, H. H. Song, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and J. Emmons. Detecting the performance impact of upgrades in large operational networks. In *Sigcomm*, 2010.
31. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1, page 156. Cambridge university press Cambridge, 2008.
32. A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, Y. Ganjali, and C. Diot. Characterization of failures in an operational IP backbone network. *IEEE/ACM ToN*, 2008.
33. V. Padmanabhan, S. Ramabhadran, S. Agarwal, and J. Padhye. A study of end-to-end web access failures. In *CoNEXT*, 2006.
34. 2013 cost of data center outages, 2013. `http://www.emersonnetworkpower.com/documentation/en-us/brands/liebert/documents/white%20papers/2013_emerson_data_center_cost_downtime_sl-24680.pdf`.
35. M. Prince. How to launch a 65Gbps DDoS , and how to stop one, 2012. `http://blog.cloudflare.com/65gbps-ddos-no-problem`.
36. J. Ramos. Using TF-IDF to determine word relevance in document queries. In *Proc. International Conference on Machine Learning (ICML)*, 2003.
37. J. J. Rocchio. Relevance feedback in information retrieval, 1971. `http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`.
38. V. Rode. Outages – outages (planned & unplanned) reporting. `https://puck.nether.net/mailman/listinfo/outages`.
39. A. Shaikh, C. Isett, A. Greenberg, M. Roughan, and J. Gottlieb. A case study of OSPF behavior in a large enterprise network. In *ACM IMW*, 2002.
40. Sophos user bulletin board. `https://www.astaro.org/gateway-products/general-discussion/44500-ddos-attack-via-dns.html`.
41. M. B. Tariq, A. Zeitoun, V. Valancius, N. Feamster, and M. Ammar. Answering "what-if" deployment and configuration questions with WISE. In *Sigcomm*, 2008.
42. K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL*, 2003.
43. D. Turner, K. Levchenko, A. C. Snoeren, and S. Savage. California fault lines: Understanding the causes and impact of network failures. In *SIGCOMM*, 2010.
44. K. Veropoulos, C. Campbell, N. Cristianini, et al. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on artificial intelligence*, volume 1999, pages 55–60. Citeseer, 1999.
45. E. Vielmetti, 2012. `http://goo.gl/ODnq5q`.
46. K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al. Constrained k-means clustering with background knowledge. In *International Conference on Machine Learning (ICML)*, volume 1, pages 577–584, 2001.
47. D. Watson, F. Jahanian, and C. Labovitz. Experiences with monitoring OSPF on a regional service provider network. In *ICDCS*, 2003.