Using Large-Scale Social Media Networks as a Scalable Sensing System for Modeling Real-Time Energy Utilization Patterns

Todd Bodnar, *Member, IEEE*, Matthew L. Dering, Conrad Tucker, *Member, IEEE*, and Kenneth M. Hopkinson, *Senior Member, IEEE*

***Abstract*—The hypothesis of this paper is that topics, expressed through large-scale social media networks, approximate electricity utilization events (e.g., using high power consumption devices such as a dryer) with high accuracy. Traditionally, researchers have proposed the use of smart meters to model device-specific electricity utilization patterns. However, these techniques suffer from scalability and cost challenges. To mitigate these challenges, we propose a social media network-driven model that utilizes large-scale textual and geospatial data to approximate electricity utilization patterns, without the need for physical hardware systems (e.g., such as smart meters), hereby providing a readily scalable source of data. The methodology is validated by considering the problem of electricity use disaggregation, where energy consumption rates from a nine-month period in San Diego, coupled with 1.8 million tweets from the same location and time span, are utilized to automatically determine activities that require large or small amounts of electricity to accomplish. The system determines 200 topics on which to detect electricity-related events and finds 38 of these to be valid descriptors of energy utilization. In addition, a comparison with electricity consumption patterns published by domain experts in the energy sector shows that our methodology both reproduces the topics reported by experts, while discovering additional topics. Finally, the generalizability of our model is compared with a weather-based model, provided by the U.S. Department of Energy.**

***Index Terms*—Event detection, Granger causality, predictive models, social network services, unsupervised learning.**

Manuscript received January 11, 2016; accepted October 3, 2016. This work was supported by the Air Force Summer Faculty Fellows Program. This paper was recommended by Associate Editor Y. Wang.

T. Bodnar is with the Center for Infectious Disease Dynamics, Pennsylvania State University, State College, PA 16802 USA (e-mail: toddbodnar@gmail.com).

M. L. Dering is with the Department of Computer Science and Engineering, Pennsylvania State University, State College, PA 16802 USA (e-mail: mld284@cse.psu.edu).

C. Tucker is with the Department of Engineering Design, Pennsylvania State University, State College, PA 16802 USA, the Department of Industrial Engineering, Pennsylvania State University, State College, PA 16802 USA, and also with the Department of Computer Science and Engineering, Pennsylvania State University, State College, PA 16802 USA (e-mail: ctucker4@psu.edu).

K. M. Hopkinson is with the Department Electrical and Computer Engineering, Air Force Institute of Technology, Dayton, OH 45433 USA (e-mail: kenneth.hopkinson@afit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2618860

I. Introduction

Social media network models have the potential to serve as dynamic, ubiquitous sensing systems that serve as an approximation of physical sensors with the added benefits of: 1) being scalable; 2) publicly available; and 3) having lower setup and maintenance cost, compared to certain physical sensors (e.g., smart meters or smart plugs). Each day, social media services such as Twitter, Facebook, and Google, process anywhere between 12 terabytes (10^{12}) [1] to 20 petabytes (10^{15}) [2] of data, making them suitable for large-scale data mining and knowledge discovery. The ability of individuals within a social media network to: 1) detect a phenomenon; 2) observe and interpret a phenomenon; and 3) report the impact of the phenomenon back to the social media network in a timely and efficient manner, highlights the potential for social media networks to be perceived as large-scale sensor networks. However, as with many large-scale sensor systems, the fundamental challenge is separating signal from noise. The conventional wisdom has been that in order to accurately understand a complex phenomenon (e.g., energy utilization patterns), complex sensors are required (e.g., smart meters) to sense, collect data, and make inferences in real time. This paper aims to challenge these conventional paradigms of social media networks and physical sensor systems by demonstrating the viability of social media networks to be used as dynamic, ubiquitous sensing systems that provide comparable level of information and knowledge, to physical sensor systems setup to achieve similar objectives.

In this paper, we propose a system that automatically generates and tests relationships between topics on social media network and electricity usage pattern. These topics are then used to predict future electricity use or test Granger causal links between the topics and the usage. This Granger causality is used to validate these links. We consider a case study where our methods are applied to energy use disaggregation using social media network data. That is, can our system discover interesting relations in social media networks that trend with electricity consumption rates? We then compare the topics that our system detects to be valid against actual topics chosen by an expert in the energy domain or against keywords mined directly from the dataset. We find that, in addition to other topics, our system replicates the topics chosen by an expert. Furthermore, a direct comparison to keyword analysis results in up to a 16.7% improvement in detected correlations (as described in Section V-B). Finally, a comparison with a weather-based simulation of homes in cities is considered.

In this paper, we provide an implementation, quantitative evaluation, and analysis of this mapping. In Section II, previous work on social media network analysis, topic modeling, and electricity use disaggregation is discussed. In Section III, a formal implementation of this mapping system is provided. In Section IV, a case study is presented where $y =$ *electricity consumption rates,* and $\mathbf{X}$ is statistically derived social media network data. In Section V, this method of hypothesis generation is compared against expert-based and machine learning-based hypothesis generation. In Section VI, we test our model's capability to predict future electricity usage. In Section VII, we conclude.

II. Previous Work

A. Mining Social Media Networks

Social media networks are emerging as the next frontier for novel information discovery. Previous work has shown applications toward measuring weather patterns [3], diagnosing illness [4], tracking earthquakes [5], providing user recommendations [6], exploring plans of action in crises [7], detecting security risks [8], and describing obesity patterns [9]. Part of social media network's advantage is the relatively openness and ease of collection of data, which, unlike traditional websites, are created by a larger population of users whose demographics are more representative of the general population [10].

One way that social media network data can be represented is as a set of sensors, where each user is a noisy sensor [4], [5]. That is, instead of reporting numerical data like traditional sensors do, social media network users report textual data which must be preprocessed before statistical methods can be applied. Simple keyword analysis—a mainstay of modern text analysis—can be problematic when applied to big datasets. For example, Google Flu Trends' system of applying text analysis to search queries has been shown to over estimate ground truth influenza rates [11], [12]. In this paper, we employ topic modeling to avoid the worst case scenario of an exhaustive search of keyword-phenomena relations.

B. Topic Modeling

Topic modeling is a way to algorithmically derive topics from unstructured documents of text. Modern work has been focused on latent Dirichlet allocation (LDA) and its derivatives [13]–[15]. LDA works by determining clusters of words in a document to determine "topics" through a Bayesian process. These topics can be represented by the words that, statistically, best describe the cluster. It has been shown that LDA can be used to detect topics in datasets such as Wikipedia articles [16], [17], scientific literature [18], spam classification [19], news analysis [20], and tweeting behavior [9], [21]. In this paper, we demonstrate that the set of topics generated by topic modeling algorithms are indeed statistically valid approximations of events. We further show that by mining these event-phenomena patterns, researchers can discover events strongly related to phenomena of interest.

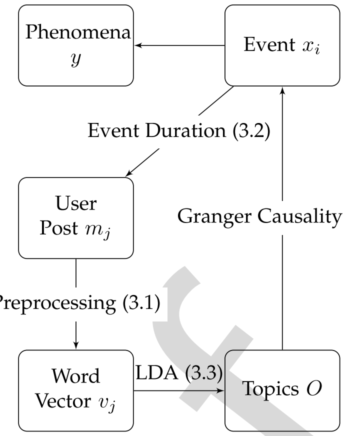

Fig. 1. High-level description of our system to transform a social media network stream into hypotheses about a real world event.

C. Knowledge Management in Energy Systems

Smart grids use communication to facilitate context awareness and cooperation across much wider areas than previous power grid system [22]. Among the initiatives to introduce smart grids are the advance metering infrastructure [23], [24] for metering in the distribution system, demand pricing, IEC 61 850 substation automation [25], the wide area management system [25], [26] for wide-area PMU measurement, and the North American synchrophasor initiative [27] that uses wide-area utility communication. Smart grid operations rely on periodic collection of data through sensors followed by processing the data.

The technology provided by a smart grid is valuable for reducing or predicting large spikes in electricity utilization. For example, by coordinating households to not perform high power usage activities concurrently. However, the smart grid has not yet been widely implemented. This paper has focused on methods to study nonsmart grid data to study either high-level usage patterns, such as total energy consumption in a city, low-level usage patterns to measure device level energy consumption, or placement of systems based on simulation [28]. It would be difficult to generalize high-level measurements to work at a finer grain because real-time electric consumption sensors are typically deployed on a station or node level. Thus analysis is limited to events that impact a large area, such as the temperature or time of day [29]–[35]. Low-level, device-based measurements have been proposed as a method to disaggregate high-level power consumption patterns [36]–[38]. These sensor networks have the advantage of providing device-level information and bypassing the need to rely on a power company for data. However, these are expensive to implement and require installation of hardware in the study participant's house, limiting the amount of data that can be collected.

To demonstrate the practicality of our system in real life applications, we consider applying our system of automated event detection to provide a novel system of energy usage disaggregation which can take high-level, publicly available power consumption records and generate valid hypotheses about behaviors that affect this consumption. For a graphical description of our methods (see Fig. 1). First, we clean textual social media network streams. Then we use LDA on the cleaned text to detect topics. These topics are then used as the basis for hypotheses about a real-world event. These topics are then tested for statistical significance. Validated hypotheses are then reported.

III. Social Media Network Electricity Utilization Methodology

In this section, we propose using large-scale social media network data as method of tracking a subset of events that are relevant to the social media network users, $\mathbf{X}$. That is, exposure to a particular event $x_i \in \mathbf{X}$ may induce a user to post a message m at time j, m_j on a social media network. Here, we assume m_j to be text-based. That is, it can be represented with a word vector v_j, derived from the raw message m_j. While it is easy for a user to map $x_i \rightarrow m_j$ (for example, "I need to do my laundry"), it may be hard to reverse this mapping, at least in a machine processable manner. Since our goal is to generate these x_i to test against phenomena, in this case: electricity usage, we must approach this mapping in an indirect fashion. Thus, we develop topic models from these word vectors where we assume a topic o is an approximation to event x_i for some i. Later, we provide an empirically tested and validated analysis of this assumption (see Section V). This allows us to map $m_j \rightarrow v_j \rightarrow o \rightarrow x_i$, effectively reversing the mapping of $x_i \rightarrow m_j$ in an unsupervised manner. Thus, we are able to formulate and validate statements of the form "x_i is related to phenomenon y" without prior knowledge about x_i.

A. Cleaning Raw Social Media Network Data

Social media network data are commonly described as extremely noisy [3], [5], [39], requiring intensive cleaning of the social media network stream as a necessary first step. We do this by converting a string of characters into a list of n-grams—pairs of up to n contiguous words (see Algorithm 1). The n-grams are determined by tokenizing the string on all nonalphabetical characters. Since capitalization can be erratic in social media networks, the n-grams are then converted to lowercase. As the objective of this step is to derive topics instead of keywords, we stem each of these words using porter stemming [40]. This maps words with similar stems but with different suffixes to the same keyword. For example, "accept," "accepting," and "acceptance" are all mapped to the same keyword, accept.

This list of n-grams is expected to follow a long-tail distribution [41], resulting in the likelihood that some are too common or too rare to be valuable in the analysis. Common words such as "the," "is," and "and" give little or no information about the text and could overshadow other, more descriptive, words that do not occur as frequently [13], [17], [42]. Thus common words, as defined by Lewis *et al.*'s [42] *stop list*, are removed from the list of n-grams. On the other hand, if a word is too rare, it may not occur enough for any inferences about it to be generalizable. Since the distribution of n-grams has a long-tail, most words will be too rare. Thus there is the potential of these very-rare n-grams to lower our ability to generate inferences about *any* n-grams [4], [17], [41]. This problem

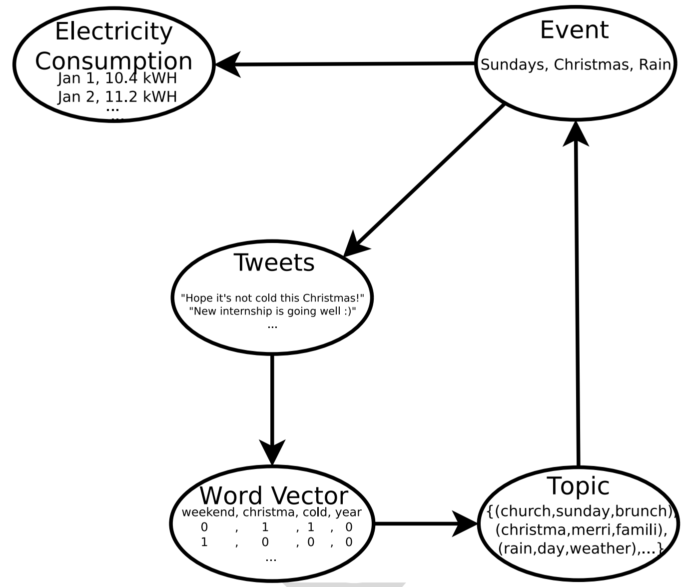

Fig. 2. Implementation of our theoretical model (see Fig. 1) for our case study.

is addressed by removing any n-grams that occur less than c times [4], [17], [20], [41], [42]. However, previous work tends to be somewhat vague on how to determine c, often incorporating expert knowledge to determine c. Here, we determine c algorithmically.

To determine c, first begin with the distribution of n-gram counts. That is, f_m is the number of n-grams that occur exactly m times each in the dataset. We then iteratively test each value for $c > 0$ until we find the minimum value for c such that

$$\frac{f_c}{\sum_{m=c+1}^{\infty} f_m} < \delta_{\min} \qquad (1)$$

where $\delta_{\min}$ is a user defined stopping threshold. Thus, we define rare words as words that occur less than δ times and remove them—a necessary step for preprocessing for LDA [17].

Note that we specifically do not remove keywords related to URLs as they may provide additional information about the user's activity. For example, tweets containing a link copied from a Web browser are likely to include "http" which may be less common on mobile users. Alternatively, links with "4sq" (reduced to "sq" when numerics are removed) are sent through four square's—a popular location check-in service—mobile application, informing us that the user is more likely visiting a location outside of his or her house.

B. Pairing Real World and Social Media Network Data

Social media network data can be updated on a millisecond level; however, it is rare for real-world events to be reported at such a temporal resolution. Additionally, it is unlikely that a single social media network message contains significant, relevant information about the real-world event we want to study, or if it does, they are exceedingly rare. We address this discrepancy by normalizing the social media network data to the real-world data's time scale. That is, we define a document

Algorithm 1: Preprocessing Steps for Social Media Network Data

Data: Time tagged Messages $\mathbf{M}$
Result: A set of aggregated and processed messages $\mathbf{D}$
d_q = document of keywords at time q;
$count_{word}$ = frequency of "word" in all documents;
$\mathbf{W}$ = set of all known stemmed words;
for $\mathbf{m_j} \in \mathbf{M}$ **do**
    Break $\mathbf{m_j}$ into substrings on non-alphabetical characters ^$[a - zA - Z]$;
    $\mathbf{j}$ = hour $\mathbf{m_j}$ was posted;
    for *non-empty Substring* $\mathbf{S}$ *in* $\mathbf{m_j}$ **do**
        convert $\mathbf{S}$ to lowercase;
        stem $\mathbf{S}$ using porter stemmer;
        add $\mathbf{S}$ to $\mathbf{W}$;
        push $\mathbf{S}$ onto $d_{\mathbf{j}}$;
        $count_{\mathbf{S}}$ ++;
    end
end
for *word* $\mathbf{S}$ *in* $\mathbf{W}$ **do**
    if $count_{\mathbf{S}} < \delta_{min}$ **then**
        Remove $\mathbf{S}$ from each d_q;
        Remove $\mathbf{S}$ from $\mathbf{W}$;
    end
end

Algorithm 2: LDA Algorithm in the Context of the Proposed Social Media Network Model

Data: set of **Documents** D, topics $\mathbf{O}$
Result: a $|\mathbf{W}| \times |\mathbf{O}|$ matrix
for Document $d \in D$ **do**
    for *Word* $w \in d$ **do**
        w_{topic} = Random topic $\in \{0, \ldots, |\mathbf{O}|\}$;
    end
end
for *Step in* $\{1, \ldots,$ *stop point*$\}$ **do**
    for Document $d \in D$ **do**
        for *word* $w \in d$ **do**
            for *topic* $o \in \{0, \ldots, |\mathbf{O}|\}$ **do**
                $P(o|d) = \dfrac{|w \in d \text{ where } w_{topic} = o|}{|w \in d|}$;
                $P(w|o) = \dfrac{|w \in D \text{ where } w_{topic} = o|}{|w \in D|}$;
            end
            Assign w_{topic} based on $P(w|o) \times P(o|d)$.
        end
    end
end

d_j to be the aggregation of all processed social media network messages v_j (as derived from m_j) that occur in during the timespan between the qth real-world event x_q and the next event, x_{q+1}. More formally

$$d_j = \left\{v_j | \text{time}\left(x_q\right) \leq \text{time}\left(m_j\right) < \text{time}\left(x_{q+1}\right)\right\} \quad (2)$$

where v_j is the n-gram representation of message m_j and time(e) is the time when e occurs. For example, if one is looking at temperature data that is reported on an hourly basis, a document would be all posts that occur within that hour. Algorithm 1 outlines how these messages are processed into word vectors, and subsequently aggregated into a document. It would be unreasonable to assume that a user posts a message *exactly* when the event happens. Instead, it is likely that the user posts about an event sometime *before*, *during*, or *after* the time that the event occurs. This issue is partially addressed when the data is aggregated, because all message after an event, but before the next, will be combined, regardless of lag between event and message.

Additionally, data can be paired based on geospatial information, such as which zip code the message occurred in. This is dependent on the dataset describing the phenomena y and the social media network messages $m_j \in \mathbf{M}$ both containing comparable location data. Caution should be advised if arbitrary spatial units are defined: the "modifiable areal unit problem" can bias results from geospatial aggregation and remains an open problem [43], [44].

C. Generating Topic Models

A given set of documents defined by the aggregation described above can be used to generate topics through LDA. We use Gibbs sampling [17], [18] implemented by JGibbLDA [17] to perform this analysis. LDA determines the probability of a document being about a topic given that it contains a set of n-grams [13], [17], [18]. To do this, LDA first generates clusters of words based on co-occurrence in the documents. That is, the probability of a word w occurring given that a document is in topic o_w. To represent these topics in a human readable form (for example, in Tables I and II), we present the set of words that have the highest probability of occurring within the topic. In other words, the topics can be expressed as a $|\mathbf{W}| \times |\mathbf{O}|$ matrix, where $\mathbf{W}$ is the vocabulary found in Section III-A and $\mathbf{O}$ are the topics generated by the LDA model such that $\mathbf{o} \in \mathbf{O}$. Each entry in this matrix corresponds to the probability of that word belonging to that topic. LDA works according to Algorithm 2. Note that the stop point is selected as 2000, the default of JGibbLDA as proposed by Heinrich [45]. This algorithm uses as input each of the aggregated Documents from Algorithm 1 to generate O topics.

The probabilities contained in this matrix can be reversed using Bayes' theorem to determine the probability that a document is in topic o given that it contains a set of keywords. Since each document has a related time component, we can say that the probability of a document being in o varies over time. By considering the likelihood of all topics over all documents, we can observe the changing interests of the population of users over time. Each of these topics are the basis of a question: "Question: Is the ith event x_i (as inferred from topic o) related to real world phenomena y?"

D. Determining Event-Phenomena Causality

In Section III-C, we outlined the method to generated topics—which we later show in Section IV-C to be statistically

Algorithm 3: Mapping Topics to Effects

Data: **Documents** D and Topics **O** from Algorithms 1 and 2
Result: Granger Causal Topics
for *document* $\mathbf{d} \in D$ **do**
  for *topic* $\mathbf{o} \in \mathbf{O}$ **do**
    $\mathbf{TS_{o,d}}$ = rate of **o** in **d**
  end
end
for $\mathbf{o} \in \mathbf{O}$ **do**
  Significance = `Granger(`$\mathbf{TS_o}$,**PowerUsage**`)`;
  if Significance then
    Print **o**;
  end
end

Algorithm 4: Computational Complexity of This Methodology

input : Social Media Posts
output: Predictions
Social Media Posts arrive: $\mathcal{O}(1)$;
Preprocessing: $\mathcal{O}(m)$ where m = number of posts;
topics ← Generate Topics (LDA): $\mathcal{O}(Nm^2)$ (see alg 2);
CausalTopics ← `Granger(`topics`)` $\mathcal{O}(Len(\text{topics}))$;

valid approximations of events—from social media network and determined the frequency of each topic at a given time. Next, we explore the patterns of each of these events over time. That is, combining all frequencies of an event over time results in a time series to be compared to the real world phenomena. Some topics, such as *Christmas*, *hating Mondays*, or *having lunch* will display cyclical patterns while other events, such as ones about a *hurricane* or a *concert*, may be one-time, anomalous events.

The event's time series can be compared to the document time series related to the real-world phenomena through cross-correlation (see Algorithm 3). That is, by matching events frequencies and real world phenomena by their time, can we find any relations between the two variables? This is defined by the Pearson's rank correlation where each point is a pairing of event frequencies and real world phenomena. The system does not filter by positive or negative correlation: a strong negative relationship between an event and a real world event can be just as interesting as a positive one. While these correlations may be strong, they do not necessarily imply a causal link.

While we do consider a correlative analysis between automatically detected events and electricity consumption, there is also an interest in determining which—if any—of the behaviors have a causal relationship on the electricity rates. Detecting strong causality through an uncontrolled, observational study without an external model of the system is impossible. Hence, we focus on detecting Granger causality [46], [47], a less stringent form of causal testing. Simply put, "correlation does not imply causality" because there may be a third phenomena that influences both, or if there is a causal relation between the two phenomenas, it is impossible to tell which one causes the other without external information. Granger causality addresses the second issue by employing lagged data. This aids in establishing a causal relationship by testing not only the synchronous variables, but measuring if the lagged data aids in the explanatory power of the model. That is, can information about phenomena y at time t (y_t) be inferred by a behavior x at time $t-t'$, for some positive value of t'? If it can, then we at least know which direction causality is flowing. To control for auto-correlative effects, the standard model compares an auto-correlation model of the predicted phenomena y

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_{(\text{lag}_{\max})} y_{(t-\text{lag}_{\max})} \tag{3}$$

where $\text{lag}_{\max}$ is the maximum lag considered in the model, determined by maximum likelihood estimation. We then add the lagged components from an event's trend x_i to the formula

$$\begin{aligned} y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_{(\text{lag}_{\max})} y_{(t-\text{lag}_{\max})} \\ + \beta_{(2+\text{lag}_{\max})} x_{i,t-1} + \ldots + \beta_{(2*\text{lag}_{\max})} x_{i,(t-\text{lag}_{\max})}. \end{aligned} \tag{4}$$

The predictive power of these two models is compared by performing a t-test on the errors between the two models. If we find that (4) performs better than (3), then it is because knowledge about this second event informs us about the future state of the target phenomena. While this is still not a test for true causality, Granger [46] have argued that it is a step in that direction. Note that Granger causality does not control for a third phenomena, which influences both the ith x in question, x_i, and y, other than guaranteeing that it occurs at some point before y. Indeed, in our case, we assume that a behavior influences both x_i—tweeting about the risk factor—and y—later power consumption due to the behavior. This method of dual time series analysis has two benefits: it quantifies how long of a lag is meaningful, and determines which sampled topics are significant.

This Granger causal test allows us to quantify the causal relationship between a phenomenon (a change in power usage) and an event (as represented by one or more social media topics). This causality measurement is the primary method of establishing causality implemented in this methodology. Social media posts can be processed into topics ahead of time, and these topics can be detected within new posts in linear time. This also allows these causal relationships to be updated in an online fashion. If the performance of the predictive nature of these causal relationships degrades, a new sample can be drawn and recalculated (see Algorithm 4). This allows us to adapt and use new data instead of relying solely on old data.

E. Validating Event-Phenomena Relationships

At this point, we have generated relationships of the form: "Topic o is related to a real world phenomena y with correlation r_o." However, if the coefficient of determination, r_o^2, is small, then any trends detected may not be statistically significant. Thus, we calculate the p-value for each regression. Since the system may test hundreds or thousands of regressions,

Algorithm 5: Topics to Predictions

Data: Significant Time Series S, **PowerUsage** Data
Result: Measurement of Predictive value of Social Media Network data
Let $\mathbf{PowerUsage}_h = \mathbf{PowerUsage}$ data lagged by h hours;
$\mathbf{S}_{a,b} = a^{th}$ significant Time Series lagged by b hours;
Build model $f(\mathbf{S}_{a,b}, \mathbf{PowerUsage}_b) = \mathbf{PowerUsage}$;
Evaluate f on data from subsequent time period;

the traditionally chosen cut off $\alpha = 0.05$ must be corrected. That is, if 100 tests are conducted on randomly generated data, it is likely that five will be reported as false positives. Bonferroni correction [48] was chosen because it does not depend on normal distribution or independence assumptions. Bonferroni correction defines the corrected cut off as $\alpha' = \alpha/n$, where n is the total number of hypotheses tested. This method of correction is more conservative than others, giving more assurance that any hypotheses that do pass the test are valid.

By implementing our system, events can be inferred from social media network data which can inform researchers about real world phenomena, as we will show in Sections IV–V. Finally, we evaluate the predictive value of this methodology as outlined in Algorithm 5 in Section VI.

IV. Case Study

In this section, we demonstrate the feasibility of our system on Twitter data, in order to determine whether topics can help explain a real world energy utilization (see Fig. 1). Specifically, we consider electricity consumption from single-family households in San Diego County from March 3, 2011 to December 31, 2011 and 1.8 million tweets from the same timespan that originated from San Diego County. That is, $\mathbf{M}$ ={Tweets in San Diego between March 3, 2011 and December 31, 2011} and y = *electricity consumption rates in kilowatts.*

A. Description of Datasets

Electricity consumption data was provided by the San Diego County Gas and Electric Company which supplies power to residents of the San Diego County in southern California. Data was provided on a daily basis for the year 2011 and represents a typical, single-family, residence.[1] Power usage data was discarded before the initial collection of Twitter data on March 3. Since power usage has both a daily cycle and longer-term dynamics (see Fig. 3), we consider both hourly and daily aggregation of the data.

Twitter data was collected between March 3, 2011 and December 31, 2011 through the Twitter API by searching for all tweets with high-resolution geospatial data. Additionally, tweets are filtered to be located within San Diego County as defined by the 2011 TIGER shape file[2] to match the spatial boundaries of the power data. A total of 1 813 689 tweets matched this criteria. The raw *jsons* returned by the Twitter API were then processed through The Open Twitter Parser[3] and stored in an MySQL database for further processing.

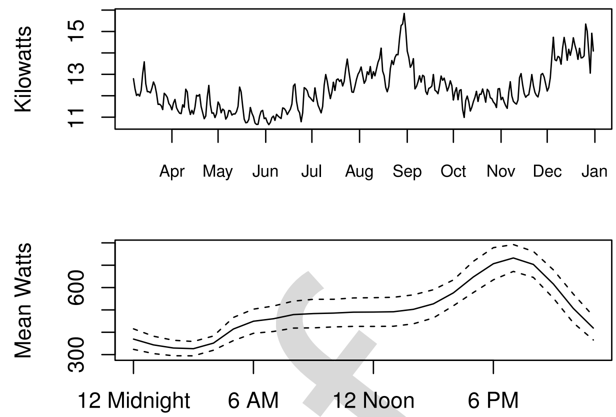

Fig. 3. Mean daily and hourly rates of power consumption for San Diego residents. Dashed lines in hourly graph indicate one standard deviation.

B. n-Gram Selection

Next, the Twitter data was cleaned through tokenization, stemming, case-normalization and *stop word* removing, as described in Section III-A. In this case study, we only consider unigrams (n-grams where $n = 1$) for analysis. Only unigrams are considered for several reasons. A higher dimensionality would cause the number of correlations to be calculated to explode. Furthermore, most implementations of LDA only consider unigrams, as topics must be latent relationships between these unigrams. Finally, given that the dataset of length constrained social media posts is studied, it is fairly common for users to discard words, which would severely limit the usefulness of n-grams for $n > 2$. A total of 794 917 unigrams were detected. We set $\delta_{\min}$ to one percent and determine that the optimal cut c to be 102, removing any unigrams that occurred less than 102 times in the dataset (see Fig. 6), thus we define $\delta_{\min}$ and use this to calculate c, which allows our approach to scale to datasize. This automated selection of c generates comparable results to other papers [16]–[18] that use domain knowledge to choose their cut off, while still allowing for more or less frequencies depending on datasize. This also helps if new samples need to be drawn and tested.

C. Knowledge Discovery of Statistically Relevant Social Media Topics

We now aim to show that these topics are statistically related to the real world events that they describe, as our assumptions in Section III require. As a null hypothesis, we consider that individuals are free to discuss any topic at any time. That is, the probability of a topic being discussed, $P(o)$ does *not* depend on the time. Instead, if our original assumption is correct, then $P(o\|x_i) > P(o)$ for some $o \in O$ and $x_i \in \mathbf{X}$. Hence, a

[1] http://www.sdge.com/sites/default/files/documents/Coastal_Single_Family_Jan_1_2011_to_Jan_1_2012.xml

[2] http://www.census.gov/geo/www/tiger/tgrshp2011/tgrshp2011.html

[3] https://github.com/ToddBodnar/Twitter-Parser

TABLE I
WORDS THAT BEST DESCRIBE THE 20 DAILY TOPICS FROM TWITTER THAT OUR SYSTEM DETERMINED TO BE ABOUT POWER USAGE AND THE CORRELATION BETWEEN THE TOPIC AND POWER USAGE. NOTE THAT THE TOPICS HAVE BEEN SORTED BY CORRELATION COEFFICIENT

r	Most likely words in the topic
-0.519	**job** http ly bit ca sandiego tinyurl **getalljob** www **tweetmyjob** lt **manag** electron soni **service** carlsbad gt ...
-0.480	sq http **instagr** la gowal ly bit job san diego tinyurl **twitpic** es lt **beach** sandiego **day great foursquar** www ...
-0.344	**jobcircl cybercod job** ca **engin develop hire softwar** mesa **sale** www **senior** la **design manag** net voic **game web** ...
-0.335	work gt dr check street **offic** fit show diego **hour** center **facebook starbuck art** airport media mesa **lunch** busi ...
-0.301	rt **coupon summer** spag june caseyanthoni es sandiego **lockerz** souther em poway doi gov earthquak ...
-0.282	rt lmao **june** tinyurl spag **marathon** jonez **job getalljob** upling samoan rock roll heat damsel untp final show ...
-0.281	**weekend** spag coupon **memori** back cri **sad hangov disapoint** kck **justinbeib** sandiego es oprah lockerz support ...
-0.247	wednesday **fat** thrusday **muscl** free bit **weight** ur loss hump **diet wine** market fan friday set eleddieg hot ...
0.201	glass **sun** auto **sprinkler** rek rt **repair** xd **replac** tcot pancak commanderlov pae coupon del word mar ...
0.211	**real pretend** jlh thereal point don itsatumblrth year iamlaceychabert laceyoffici **handbag design manufactur** ...
0.225	jlh **frenchfan** victoria **witter clalovehewitt** lol alexandria thereal don rt tweet es **coupon** ya bcuz camill game ...
0.225	**christma cold** year dat jus sir final ass wyd bro si lo nba yea smh man dnt crystal twitter laker victoria ...
0.238	yummi day **sexi orgasm** good morn email hotmail saturn great **love beauti** school video lick class cum pretti ...
0.254	**christma merri famili eve xmas happi holiday santa** present lt **gift** year stephazilla laker **church navidad** hous ...
0.267	de eu pra um na da se vou mas uma mai meu tem vai em ver happi dia por ele minha person didn beauti ...
0.297	http san diego love **good time** day don make today back la job ca people haha lol **home** feel ll wait great ...
0.369	http **vista shop plaza valley** chula center **mall fashion buy** bonita peopl mission pkwi **store** home break work...
0.410	lt gt **lol fuck shit** job haha ca dr **ass bitch** don sandiego nigga **hate** girl feel love sleep drink ave **damn** ...
0.418	rek beauti **window hurrican** coupon iren xd vma video arhhhhjay lt omg **storm hot wind** gaga kk issu **humid** ...
0.448	lol **christma** final **holiday** dannyboyo partic travcb **home** laker deniseexclus **xmas** studi **happi** andruee ll ...

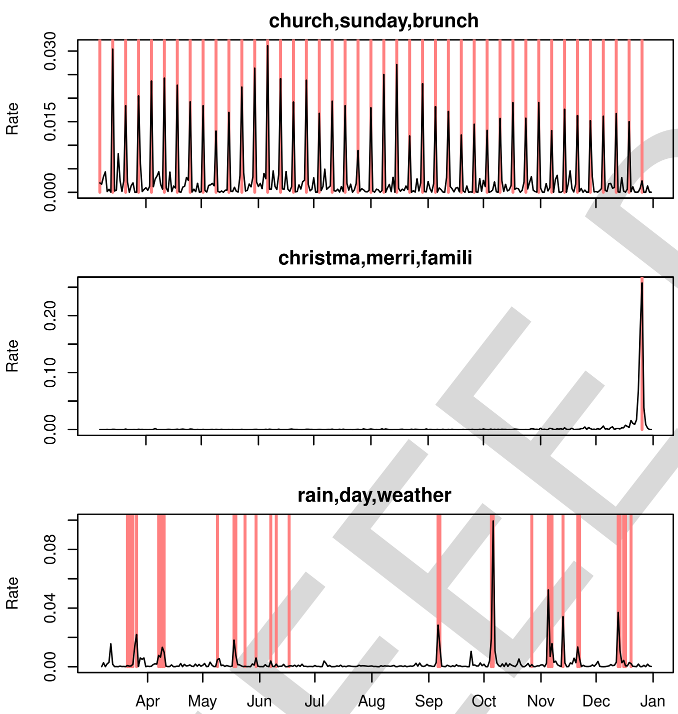

Fig. 4. Temporal patterns for three select topics in black. Chart titles indicate the three most representative words for each topic. Red lines indicate days that are Sundays, Christmas day, and days when it rained, respectively.

topic is being induced by a real world event. To verify this, we would need ground truth data for x_i, which is not necessarily possible to obtain for all possible events i. However, some topics lend themselves to easy validation.

Here, we consider three topics that appear to represent Sundays, Christmas, and rain (see Fig. 4). Sundays were chosen as the first topic for analysis because it was expected to follow clearly defined temporal patterns. Additionally, Sundays are more discrete than Christmas (i.e., are Christmas Eve and Christmas separate events?) or rain (i.e., how much precipitation is necessary for it to be considered raining?). Christmas was chosen as a topic because it is representative of events that occur only once in our dataset, but with a well defined event time. Indeed, Christmas may be the biggest topic detected, with 25.8% of the Twitter data being about Christmas on Christmas, December 25, and 17.5% on Christmas Eve, December 24. Finally, we considered rain because it lacks the periodicity of the other two topics. Note that the rate of precipitation does not have a strong relationship to spikes in the rain topic, so we discretized weather into days without rain and days with rain, as defined by weather underground.[4]

We can thus calculate the relevant probabilities (see Table I). This means that 80 topics whose correlations are too low are not present in this table. For example, with the topic sunday: $P(o_{\text{sunday}}) = 0.00350$ (as determined by LDA) and $P(o_{\text{sunday}}|x_{\text{sunday}}) = (o_{\text{sunday}} \& x_{\text{sunday}}/x_{\text{sunday}}) = 0.0182$. For completeness, we can use Bayes theorem to determine the probability that it is Sunday given that the topic is about Sunday

$$\begin{aligned} P\big(x_{\text{sunday}}|\,o_{\text{sunday}}\big) &= \frac{P\big(o_{\text{sunday}}|\,x_{\text{sunday}}\big)p\big(x_{\text{sunday}}\big)}{P\big(o_{\text{sunday}}\big)} \\ &= \frac{0.0182 * 0.15}{0.00350} = 0.728. \end{aligned} \tag{5}$$

Since we know what days we sampled from, we know that $P(x_{\text{sunday}}) = (x_{\text{sunday}}/x_{\text{all}}) = 0.14$, which is close to the general occurrences of Sundays, (one out of seven days each week $\approx$0.1429). We find that $p(\text{Event}|\text{Topic})$ is significantly higher than the baseline $P(\text{Event})$, giving evidence toward these automatically generated topics, $o \in O$ having some relation to real world events $x_i \in \mathbf{X}$.

[4]http://www.wunderground.com/history/airport/KSAN/2011/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2011&req_city=NA&req_state=NA&req_statename=NA

AQ2

TABLE II
WORDS THAT ARE MOST ASSOCIATED WITH THE 38 HOURLY TOPICS FROM TWITTER THAT DESCRIBE EVENTS THAT ARE FOUND TO GRANGER CAUSE CHANGES IN POWER USAGE

r	Most likely words in the topic
-0.432	**job** http sandiego electron ca soni sd **sonyjob tweetmyjob engin** snei **softwar director** test alskks **develop administr** . . .
-0.321	lol shit yo lmao **work** man ass good nigga dat fuck smh tat ya feel dnt de jus bitch bro **sleep** je wit home est sir tha yea . . .
-0.145	**watch movi show** love time lol good great ll fun **tv** back night yeah make year peopl awesom **episod youtub** wait tweet . . .
-0.120	job http ca sandiego kaiser **nurs tweetmyjob healthcar** permanent san diego rn kindr **hospit** ii amn **kinderedjob** account . . .
-0.106	http esriuc love lol **harri** ddlovato rt **potter** time fstk googl good day kooldudestillo pride **watch** diego rhenderson demi girl . . .
-0.086	http rt **shop** lol great san www diego **ad sale** love lmao watch june item daili don day back mile summer **inventori** time good . . .
-0.079	http **california southern earthquak** gov km usg doi june **depth** usa diego gmt hour ca mi ll good time hand join monday. . .
-0.070	http el la ma love day al ya ben de ne play ana ve wait ha lol shit da good hey ni bi man check home ik en ba wo in tweet . . .
-0.057	**lol haha love** stephazilla **good** lt **hahaha** time watch don yeah fuck night feel back thing shit girl life wait tomorrow . . .
-0.057	**victoria witter** alexandria **teamjlh** stillo http **clalovehewitt** lol stellix don back good yeah tweet **beutyqueen** gonna . . .
-0.041	http del diego san mar la **fair beach blvd counti** day school jimmi ca **camino** de pic coronado vall durant time . . .
-0.035	http **japan** www greeney san **fukushima** good rt time **nuclear** ur win day **tsunami** great plixi ipad watch diego bit . . .
-0.029	http **plaza** diego san el citi bonita horton **shop** nation **westfield** hlbd cajon ave la parkway camino time de **mall** dr **buy** . . .
-0.027	**charger** http **game** diego san qualcomm **footbal statium win play raider** good team **watch** fan **nfl** time river **tebow** rt sunday . . .
-0.021	http diego san **coronado beach hotel** mission **bay** st pic pine del torrey **resort** la ave time spa park foursquar blvd vista . . .
-0.011	**work** make today rt **offic** ll **busi** free deal market don great health week **stori** peopl year **school** citi pay list design site news . . .
-0.003	jlh thereal frenchfan love real **jennif clalovehewitt** verifi lol **hewitt** http lt fake account don tweet back good day camill make . . .
-0.003	http **day lol love** diego back don time san ca good final ll class **cold** make **break** fuck work **night** week hate haha xoxo uni . . .
0.004	np love **song** shit make fuck don back real peopl **good music** lil man girl show **listen** thing yeah **play** damn haha rt . . .
0.006	**job getalljob** ca tinyurl sandiego http engin **edit manag telecommut concierg clinic assic** sleep hotel **remot develop web** hour . . .
0.018	na ko sa hahaha haha mo ako ng ang ka lang pa naman time eh day lol ba nga good si ni oo hehe hahahaha tweet . . .
0.023	**sleep night bed goodnight tomorrow** fuck good **dream** time **tonight** wake home **asleep** hour feel love drunk sweet happi . . .
0.027	job http ca general ga poway asi atom sys **aeronaut account sandiego tweetmyjob manufactur analysis** ii iii bit **financi control** . . .
0.029	te si de la ya tu mi el esta yo como en por lo se es para mas mero hola bien con bueno muy dia una todo ke los saludo pue . . .
0.048	de http la enl en los mexico se al del lol es para funal fuck love lt work por con su son home man tv mas twitter ha una las . . .
0.077	http juli happi day don **cassey caseyanthoni** good miss san make **sagesummit** time firework ll beach life peopl bit . . .
0.081	**game rt laker** lol **win** http heat **play watch team nba** fan love final good fuck lt day **season** bull **player** ve tonight time **kobe** . . .
0.089	http **life ratio live** tune proof net diego **fit good** back time tomorrow html **work** love guy night em cujo st lol miss watch . . .
0.100	rt http time ya love lol day teamfollowback di famili yg make **cricket** ll **wireless** ur gt good haha followback yo cool . . .
0.143	http **obama dead** diego **bin** san good **war** love news time **presid laden** rt **kill** day **cnn osama** de stop vote happi . . .
0.151	http san diego lunch st ave dr pic **cafe** blvd **grill food mexican** day **burger** today mayor **taco** work foursquar offic . . .
0.172	**iphon appl steve job** app rt live http today don rip twitter wait work feel **phone die** tattoo life love **ipad** world yeah io tweet . . .
0.175	**sq instagr** gowal la ly bit **twitpic foursquar** untp **mayor beach** trendsmap street lockerz tinyurl www btw picplz year . . .
0.184	http **today morn breakfast** san **church** diego **day cafe coffe** night **sunday good starbuck** park st hour mayor pic . . .
0.195	http **morn** san diego day **good today starbuck school earli work coffe** st fit oceansid carlsbad happi blvd wake mesa . . .
0.210	http san diego st park ave fan street south experi **hotel** tomorrow **intern** year ca gaslamp fun ll **rememb** market space . . .
0.226	http san diego st washington ave **chicago** btwn street el **game pizzeria pizza** map blvd fort good **cajon** lefti . . .
0.639	san diego http **airport** intern dr **termin harbor back** work **home** hour **flight** fit earli head line great **gate miss** begin . . .

Additionally, some events will show cyclical, daily patterns (see Fig. 5). If the target phenomena also shows similar patterns, these hourly events may further help to describe the phenomena.

D. Event-Electricity Usage Relationships Detected

These automatically determined topics were found to correlate with daily power consumption rates with $-0.519 < r_i < 0.448$ (see Table I). The topic that correlated most negatively with power consumption included unigrams such as "job," "getalljob," and "tweetmyjob." This leads to the first steps of a domain expert investigating that people use less energy at their residence on days when they are at work than days when they are not working. The topic that correlated most positively with power consumption included Levins stemmed unigrams such as "christma," "holiday," and "home," hinting that people consume more electricity around Christmas time. Similarly, the topics that were determined to Granger cause changes in hourly electricity consumption correlated with the current electricity consumption between $-0.432 < r_i < 0.639$ (see Table II). As with daily rates, the topic that Granger caused the most decrease in power included unigrams such as tweetmyjob and "sonyjob."

E. Validation Steps

With Bonferroni correction for multiple tests, we determined the corrected value for $\alpha = 0.05$ to be $\alpha' = \alpha/100 = 0.0005$. Twenty correlations are found to be significant at this rate (see Table II). While we cannot make any explicit claims about the topics this citation [13] determined to have significant relations to power usage, it has been argued [9], [13], [17], [18] that the most common words in a topic are representative of the inherit meaning of the topic. Here, we present the most significant words for each topic, with select words bolded for easier interpretation. With this interpretation in mind, it appears that the three most negatively correlated topics include activity such as having a job, posting on Foursquare or Instagram (i.e., things done outside the residence) and job searches. The top three positively correlated topics include topics about Christmas, storms, and surprisingly, a topic consisting of several vulgarities.

We found a total of 20 statistically significant correlations between events (as inferred by detected topics) and power consumption. Earlier, we presented the 20 topics that had statistically significant correlations with power consumption (see Table II). However, it is also important to consider topics that are rated with a low coefficient of determination to see if

TABLE III
PROBABILITY OF A TOPIC INDEPENDENT AND DEPENDENT ON A POTENTIALLY RELATED EVENT

Topic	$p(Topic)$	$p(Topic\|Event)$	$p(Event\|Topic)$
Sunday	0.00350	0.0182	0.728
Christmas	0.00243	0.256	0.351
Rain	0.0024	0.0137	0.627

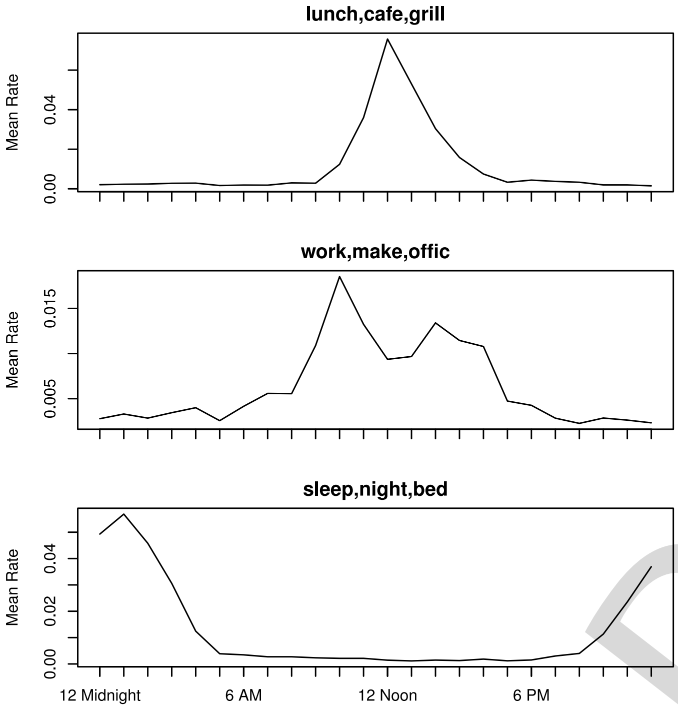

Fig. 5. Mean hourly rate of three select topics. Chart titles indicate three representative words for each topic.

TABLE IV
TOPICS GENERATED THROUGH A REVIEW OF THE LITERATURE, RANKED BY OCCURRENCE IN "NEW & USA" PAPERS

Topic	New & in USA	New	USA
Temperature	4	6	5
Income	3	4	4
Electric Price	3	4	4
Air Conditioner	2	4	5
Heater	2	2	5
Dishwasher	1	2	4
Clothes Dryer	1	2	4
Refrigerator	1	1	2
Water Heater	1	1	3
Building Codes	1	1	1
Own Pool	1	1	1
Own Spa	1	1	1
Lighting	1	1	1
Stove	0	0	3
Freezer	0	0	3
Television	0	0	2
Clothes washer	0	0	1
Wind	0	2	1
Rain	0	1	0
Household Size	0	1	0
Total Papers	7	10	10

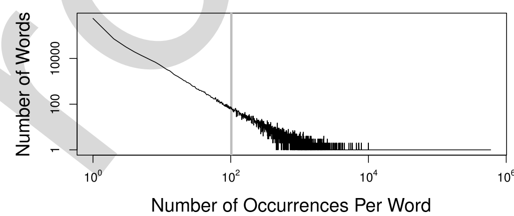

Fig. 6. Distribution of unigrams detected shows a long-tail distribution. The gray line represents the automatically determined cut, *w*.

they are actually *not* likely to related to residential electricity consumption. The least related topic's three most representative words are "asiathegreat," "manufactur" and "deal." It would appear that these topics are about manufacturing–perhaps in China–which does not have a direct effect on *residential* electricity consumption. The second least related topic's three most representative words are "louisseandon," "ya," and "blo." The third least related contains "justinbieb," "lt," and "sagesummit." These two topics would seem to be related to news about entertainers Louis Sean Don and Justin Bieber, which are likely related to entertainment news rather than electricity consumption.

V. EXPERIMENTS AND RESULTS

One may ask "what is the value of this system over traditional keyword mining or just using expert knowledge?" While our system allows knowledge discovery with limited need for expert knowledge, if it does not perform well, then it is not useful. To justify our system's existence, we compare the results of our system to topics common in the power consumption literature. Additionally, we perform keyword mining to detect words, instead of topics, that are related to electricity consumption.

A. Comparison to Domain Experts

To approximate that knowledge of an expert on power consumption modeling, we perform a literature review. We sample Google Scholar for 100 papers that appear relevant to our question. We discard 85 papers which are either inaccessible (e.g., out of print papers from the '70s), irrelevant to our topic (e.g., a paper on building the Nigerian power grid) or do not explicitly state activities to model (e.g., a paper on synchronizing houses on a smart grid which filter out the customers activities). While we could read the papers for other ideas of important topics, we avoid to because: 1) we risk biasing the set of topics due to selective reading; 2) if a topic is not explicitly modeled or measured, we can assume that the expert does not consider it important; and 3) this literature review is not designed to collect all relevant topics, just ones that are common amongst experts.

Additionally, we separate papers that are more than 10 years old or do not focus on American populations. While these papers may contain expert knowledge, our Twitter and power datasets are based on recent, American usage, which may be different from older usage patterns or those of citizens of other countries. In total, we find 12 topics from recent and local papers [30], [31], [33], [34], [49]–[51] and an additional eight topics from other papers [32], [35], [52]–[57] (see Table IV). Topics were explicitly presented from the papers by either tables or equations. If we only consider the topics that occur more than once in the set of recent and local papers ("temperature," "income," "electricity price," "air conditioner," and "heater"), then we can informally detect two clusters of topics: 1) "climate control" and 2) "economic factors." Both of these two topics were also discovered to be significant measures of electric consumption through our automated system.

Our system found 20 topics that are related to electricity consumption. Our literature review also found 20 topics that are related to electricity consumption. It would seem, however, that these two methods of knowledge discovery discovered topics that were different from each other. The literature review found topics such as temperature or dishwasher usage as interesting topics (see Table IV) while the topic modeling found topics such as having a hangover on the weekend or going to the mall as interesting topics (see Table I). This can be explained by the methods used to collect data. The literature focuses on things that are easy to measure by traditional sensors. However, we use humans as "organic" sensors. This results in different types of data collected: it is easy to have a person report that they are going out on the weekend, but relatively hard to design a sensor to measure this. On the other hand, a sensor to measure temperature is trivial to acquire, but it is unlikely for a person to accurately report the temperature on a regular basis. By focusing on the human element, we have been able to detect important factors of electricity consumption that were previously overlooked due to limitations in traditional sensors and domain knowledge.

Often times, the elements which can easily be studied by these experts and events which are present on social media do not have many commonalities. Discovering these latent events, processed by human sensors, is one major advantage of this paper over traditional sensors. For example, humans might aid in discovering a third variable at work (such as a football game), which leads to an increase in power consumption, while a more guided approach will tend to be informed instead by a television. This demonstrates that not only can we reproduce previous results, but we can also generate novel hypotheses, as told by human sensors.

B. Comparison to Keyword Analysis

We also consider algorithmically generating keywords instead of topics. First the text is cleaned through stemming and *stop word* removal, equivalent to the methods implemented in our system (see Section III-A). Instead of using topic modeling to filter out irrelevant keywords, we are limited to just selecting keywords based on their frequency in the dataset. The $n = 1, 2, \ldots, 5000$ most commonly occurring keywords are selected. The keywords are then tested for relations through cross correlation with the electricity consumption data, the same way that topics were tested for relations in Sections III-D and III-E. We try different values of n because if we try too few keywords, important keywords will be lost, but if we try too many keywords, then, once Bonferroni correction is applied, there will not be enough statistical power to detect significant keywords.

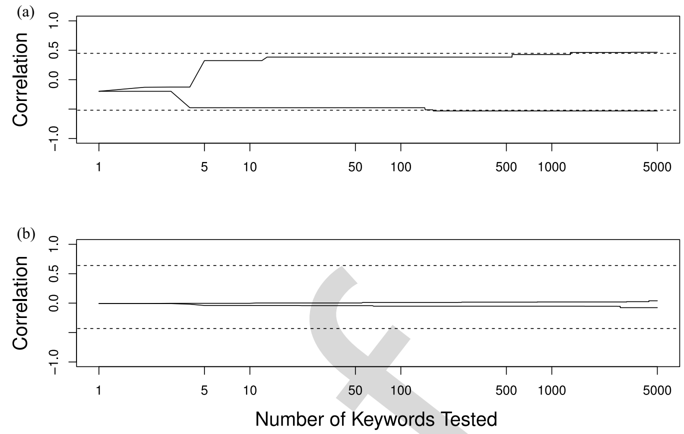

Fig. 7. Strongest positive or negative keyword given a set number of keywords tested. Dashed lines indicate the strongest positive or negative topic detected. Data was aggregated by (a) day or (b) hour.

Additionally, we could define words that occur very frequently in our dataset as de-facto stop words and remove them in addition to the predefined stop word list. However, we do not do this as the tests in this section are independent of each other (besides the Bonferroni correction), compared to the frequency-based methods of our proposed event inference system, so the gain in statistical power is limited in comparison of the risk of removing strongly predictive keywords. Finally, we consider the strongest positive and negative rates of correlation detected for each value of n (see Fig. 7). All minimum and maximum correlations displayed are significant at the 0.05 level, even when Bonferroni correction is applied.

Testing keywords instead of topics resulted in some correlations when dealing with daily aggregation. However, our keyword test allows for a number of tests equivalent to the size of the corpus, which is hard to directly compare against testing 100 topics. When we only consider the top 100 keywords, we find keywords with the strongest positive correlation to be "don" with $r = 0.384$ and the keywords with the strongest negative correlation to be sq with $r = -0.476$. Our system finds events where the strongest positive correlation is 0.448 and the strongest negative correlation of -0.519, a 16.7% and 9.03% improvement, respectively. While keyword-based models do provide some information for daily prediction, hourly prediction does not seem well suited for keyword analysis with correlations ranging between -0.074 and 0.004, limiting the usefulness of previous methods for fine-grained prediction. Comparatively, our system which finds topics that match power usage with correlations between -0.432 and 0.639 resulting in an increase of explained variance of up to 41%.

VI. Predicting Future Electrical Consumption

Up to this point we have only considered individual topics to predict the phenomena. Here, we consider multivariable regression based on lagged predictive variables to predict hourly power usage (see Algorithm 5). As a baseline, we consider a 12-variable auto-correlation model where the maximum lag of 12 was determined through maximum likelihood estimation. We then compare this model to

TABLE V
CORRELATION COEFFICIENTS FOR MODELS USING AUTO-CORRELATION, TOPICS, OR A SUBSET OF ATTRIBUTES

	Auto-Corr	Topics	Auto-Coor + Topics	Subset
Training Set	0.9515	0.9430	0.9788	0.9777
5-fold CV	0.9510	0.9116	0.9670	0.9682
80%/20%	0.9313	0.7152	0.9003	0.9632

TABLE VI
ROOT MEAN SQUARE ERRORS FOR MODELS USING AUTO-CORRELATION, TOPICS, OR A SUBSET OF ATTRIBUTES

	Auto-Corr	Topics	Auto-Coor + Topics	Subset
Training Set	39.6508	42.9102	26.3846	27.0747
5-fold CV	39.8758	53.2473	32.8872	32.2713
80%/20%	51.7108	121.166	66.3104	34.9691

three models: a multivariable regression on the detected topics, a multivariable regression on the 38 topics that were found to have a Granger causal relationship to electricity consumption *and* the auto-correlation model, and the second model with a subset of the attributes used. Which attributes are retained in the third model are selected through removing attributes with the smallest coefficients and refitting the model until AIC no longer improves.

We now determine the accuracy of each model by determining the correlation coefficient for either through traditional statistical methods, fivefold cross validation, or a 80%/20% test-train split. The 80%/20% test-train split is performed on data that is ordered by time where the fivefold cross validation is performed on randomly ordered data. We find that at least one of our models out perform the base-line in all three evaluation methods. Importantly, the 80%/20% test-train split represents the most realistic case of predicting future electricity usage, and our model provides an additional 4.28% explanation of electricity usage. These results can be seen in Tables V and VI.

A. Comparison With U.S. DOE Model

The U.S. Department of Energy provides Commercial and Residential hourly load profiles for typical meteorological year (TMY3) locations around the United States. These simulated values are derived from a combination of weather data from the National Solar Radiation Database,[5] regional climate-specific information (cold/very cold, hot-dry/mixed-dry, hot-humid, marine, and mixed-humid), and load profile type (high, base, and low) which define physical building characteristics such as home size, layout, insulation type, heating fuel source, and occupants. These simulations take into account very detailed electricity demands, (e.g., heat output by showers and dishwasher temperature point) and provide an hourly demand of an average household in each of hundreds of sites around the United States. Incorporating all of this information, this model presents a year-agnostic estimation of the hourly electricity usage of households across the country. That is, the model does not differentiate between 1 A.M., January 1, 2011, and 1 A.M. January 1, 2012. Rather, it assumes each hour is the same. The DOE has made this model

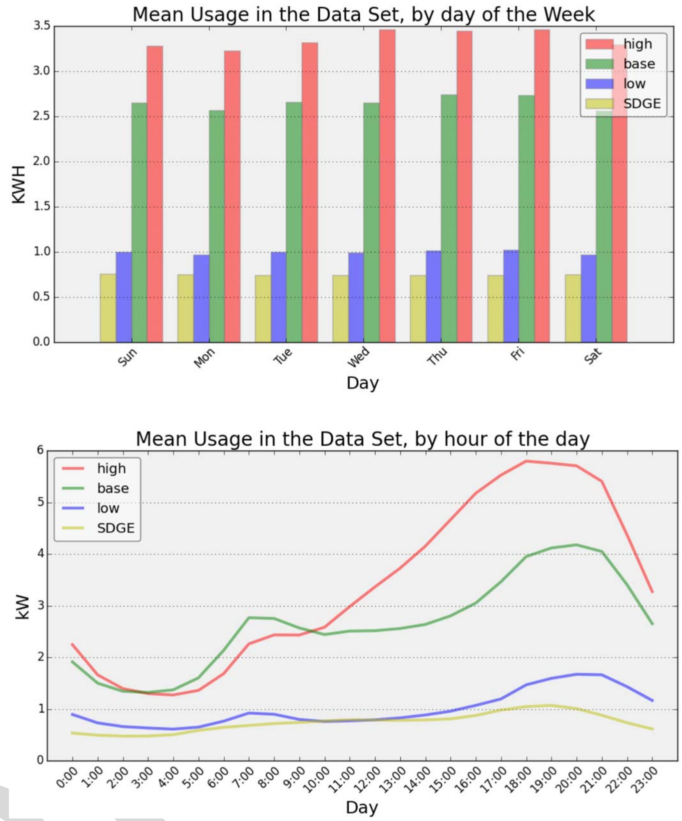

Fig. 8. Periodicity of SDGE provided energy data, compared to TMY3 simulated data.

publicly available for researchers seeking to predict energy demands across U.S. Cities.[6]

To test the efficacy of the TMY3 models in simulating the real world energy use of the San Diego area, we compared the TMY hourly use with the SDGE-provided data from Section IV. The TMY3 data is considered the baseline model, with the SDGE data representing the ground truth. Since the TMY3 data is year agnostic, variations in energy use due to severe weather events (as opposed to seasonality), and date-specific periodicity (weekends and weekdays) will not be included. These differences can be seen in Fig. 8. While the SDGE data is lower in magnitude than the TMY3 load profiles, the general trends of the data are reflected best by the *base* model, which carries an hourly correlation coefficient of 0.7544 and an RMSE of 130 when used as input for a linear regression of the SDGE data.

Next, TMY3 data is used to predict monthly SDGE electricity usage. The monthly usage data is provided by SDGE, aggregated across customers in each zip code.[7] This data is shown in Fig. 9. Note that since the TMY3 is year agnostic, the data will repeat on an annual cycle. Once again, the magnitude of each of the load models is higher than the aggregate data provided. When analyzed against the real monthly data for San Diego homes, no single model consistently correlates better than the others, with the *high* model performing best

[5]https://mapsbeta.nrel.gov/nsrdb-viewer

[6]http://en.openei.org/datasets/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states

[7]https://energydata.sdge.com/

TABLE VII
ρ AND RMSE FOR EACH TMY MODEL

year	ρ			RMSE		
	high	base	low	high	base	low
2012	0.65	0.21	0.59	121.4	156.3	129.3
2013	0.58	0.81	0.79	63.6	45.5	47.5
2014	0.82	0.78	0.93	40.7	45.1	27.2
Aggregated	0.61	0.43	0.64	83.5	94.8	80.1

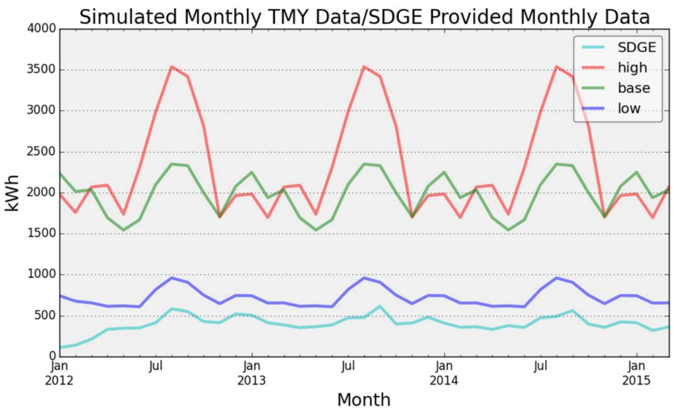

Fig. 9. TMY3 data, aggregated by month, compared with SDGE monthly data.

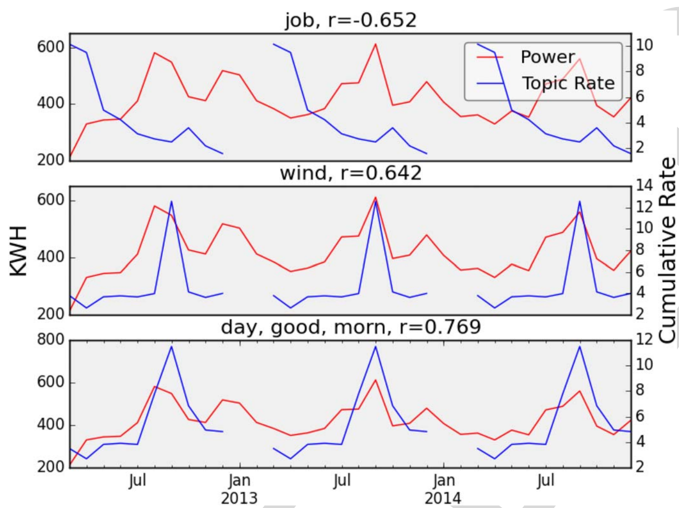

Fig. 10. Topic rates for three sample topics. Note the recurrence of the topic rate, as the topics were analyzed for 1 year only.

in 2012, *base* in 2013, and *low* in 2014. These same models possess the lowest RMSE on a yearly basis, as seen in Table VII.

Finally, we demonstrate that our proposed social media model outperforms the TMY3 model, given the same ground truth (SDGE data), by using the topic models and frequencies from Sections IV–VI. As with the TMY3 data, we assumed that each topic frequency is repeated for that same hour and date on all subsequent years. Similar to Fig. 4, these cumulative topic rates by month can be seen in Fig. 10. Next, these topics were aggregated on a monthly basis, the significance of each topic was tested, and the Bonferroni correction applied, leaving 13 topics whose $p < 0.05/100$. Finally, we used these frequencies as input in a regression model for March–December of each year. This model yielded an RMSE of 43.6 when applied to this time period, which outperforms the linear regression performance of the best TMY3 data in Table VII, whose best models RMSE was 80.1, an 83%.

VII. CONCLUSION

In this paper, we proposed a theoretical backing to our design (see Section III), which assumed a link between: 1) events and text; 2) text and word vectors; 3) word vectors and topics; 4) topics and events; and 5) events and real-world phenomena. We now provide evidence of these relations. Previous work [9], [39] has verified that events cause users to post on social media networks. Similarly, the conversion of text into word vectors has previously been discussed [4], [17], [20], [41], [42]. The most likely words are cohesive within each topic and have large between-topic variation (see Table I). Thus it is likely that topics can be generated from social media network text using LDA [14], [15]. We choose three topics that contain words related to Sundays, Christmas, and storms. By studying the temporal patterns of each topic, we find a relationship between the storm topic and the days with "rain" events in San Diego, the Sunday topic to be most often discussed on Sundays, and the Christmas topic to trend during December (see Fig. 4). Finally, we show a relationship between our discovered events and energy consumption through statistical analysis (see Table II). Hence, we conclude that there is evidence for our assumptions on links, at least when applied to our case study.

We presented a novel form of semi-supervised knowledge discovery that infers events from topics generated from social media network data. These events are then used to form hypotheses about real-world phenomena which are then validated. To provide support for our case, we perform a case study where Twitter data is used to predict electricity consumption rates. The results are then compared to topics generated by domain experts and keyword analysis. We find that our system detects events tangential to what the literature is currently focused on and that our system outperforms an equivalent keyword analysis by up to 16.7%. When combined with time-series modeling, we are able to predict electricity consumption with correlations of up to 0.9788 and a mean absolute error of 19.84 watts—less than the energy consumption of a single light bulb. Finally, we compared the performance of this model to the models generated by the DOE for the San Diego area, and found it to be more accurate.

Future work may consider a more robust comparison of this model against other existing models, since several such models exist. Additionally, this model might be employed for a more directed event detection, as described in the introduction. The textual analysis in this paper could be augmented by considering synonyms and related concepts through word embedding which groups similar words together automatically. Additionally, other data modalities might also be considered, such as images, videos, and social media metadata. Since there is a spatial component of this data, future work may also analyze similar data for a different part of the country, to determine if the trends we have identified hold true elsewhere. Finally, it may prove fruitful to analyze a similar methodology for other utilities such as water.

ACKNOWLEDGMENT

Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors. The views in this document are those of the authors and do not reflect the official policy or position of the U.S. Air Force, Department of Defense, or the U.S. Government.

REFERENCES

[1] L. Zhao, S. Sakr, A. Liu, and A. Bouguettaya, *Cloud Data Management*. New York, NY, USA: Springer 2014.

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding Twitter data with TweetXplorer," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Chicago, IL, USA, 2013, pp. 1482–1485. [Online]. Available: http://doi.acm.org/10.1145/2487575.2487703

[4] T. Bodnar, V. C. Barclay, N. Ram, C. S. Tucker, and M. Salathé, "On the ground validation of Online diagnosis with Twitter and medical records," in *Proc. 23rd Int. Conf. World Wide Web Companion*, Seoul, South Korea, 2014, pp. 651–656.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[6] M. Eirinaki, M. D. Louta, and I. Varlamis, "A trust-aware system for personalized user recommendations in social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 4, pp. 409–421, Apr. 2014.

[7] M. J. Lanham, G. P. Morgan, and K. M. Carley, "Social network modeling and agent-based simulation in support of crisis de-escalation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 1, pp. 103–110, Jan. 2014.

[8] T. Bodnar, C. Tucker, K. Hopkinson, and S. G. Bilén, "Increasing the veracity of event detection on social media networks through user trust modeling," in *Proc. IEEE Big Data*, Washington, DC, USA, 2014, pp. 636–643.

[9] D. D. Ghosh and R. Guha, "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system," *Cartography Geogr. Inf. Sci.*, vol. 40, no. 2, pp. 90–102, 2013.

[10] A. Smith and J. Brenner, *Twitter Use 2012*. Pew Internet & Amer. Life Project, 2012.

[11] T. Bodnar and M. Salathé, "Validating models for disease detection using Twitter," in *Proc. 22nd Int. Conf. World Wide Web Companion*, Rio de Janeiro, Brazil, 2013, pp. 699–702.

[12] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales," *PLoS Comput. Biol.*, vol. 9, no. 10, 2013, Art. no. e1003256.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[14] H. D. Kim *et al.*, "Mining causal topics in text data: Iterative topic modeling with time series feedback," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 885–890.

[15] X. W. Zhao, J. Wang, Y. He, J.-Y. Nie, and X. Li, "Originator or propagator?: Incorporating social role theory into topic models for twitter content analysis," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 1649–1654.

[16] M. Wahabzada, K. Kersting, A. Pilz, and C. Bauckhage, "More influence means less work: Fast latent Dirichlet allocation by influence scheduling," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, Glasgow, U.K., 2011, pp. 2273–2276.

[17] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & Web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, Beijing, China, 2008, pp. 91–100.

[18] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5228–5235, Apr. 2004.

[19] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent Dirichlet allocation in Web spam filtering," in *Proc. 4th Int. Workshop Adversarial Inf. Retrieval Web*, Beijing, China, 2008, pp. 29–32.

[20] J.-C. Guo, B.-L. Lu, Z. Li, and L. Zhang, "Logisticlda: Regularizing latent Dirichlet allocation by logistic regression," in *Proc. PACLIC*, Hong Kong, 2009, pp. 160–169.

[21] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram, "Discovering health-related knowledge in social media using ensembles of heterogeneous features," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 1685–1690.

[22] L. Dannecker *et al.*, "pEDM: Online-forecasting for smart energy analytics," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 2411–2416.

[23] V. Aravinthan, V. Namboodiri, S. Sunku, and W. Jewell, "Wireless AMI application and security for controlled home area networks," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Detroit, MI, USA, Jul. 2011, pp. 1–8.

[24] C. Bennett and D. Highfill, "Networking AMI smart meters," in *Proc. IEEE Energy 2030 Conf. ENERGY*, Atlanta, GA, USA, Nov. 2008, pp. 1–8.

[25] C. Bennett and S. B. Wicker, "Decreased time delay and security enhancement recommendations for AMI smart meter networks," in *Proc. Innov. Smart Grid Technol. (ISGT)*, Gaithersburg, MD, USA, Jan. 2010, pp. 1–6.

[26] J. E. Fadul, K. M. Hopkinson, T. R. Andel, and C. A. Sheffield, "A trust-management toolkit for smart-grid protection systems," *IEEE Trans. Power Del.*, vol. 29, no. 4, pp. 1768–1779, Aug. 2014.

[27] M. T. O. Amanullah, A. Kalam, and A. Zayegh, "Network security vulnerabilities in SCADA and EMS," in *Proc. IEEE/PES Transm. Distrib. Conf. Exhibit. Asia Pac.*, Dalian, China, 2005, pp. 1–6.

[28] P. Palensky, E. Widl, and A. Elsheikh, "Simulating cyber-physical energy systems: Challenges, tools and methods," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 3, pp. 318–326, Mar. 2014.

[29] A. Aroonruengsawat, M. Auffhammer, and A. H. Sanstad, "The impact of state level building codes on residential electricity consumption," *Energy J.*, vol. 33, no. 1, pp. 31–52, 2012.

[30] I. Ayres, S. Raseman, and A. Shih, "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage," *J. Law Econ. Org.*, vol. 29, no. 5, pp. 992–1022, 2013.

[31] I. M. L. Azevedo, M. G. Morgan, and L. Lave, "Residential and regional electricity consumption in the U.S. and EU: How much will higher prices reduce CO_2 emissions?" *Elect. J.*, vol. 24, no. 1, pp. 21–29, 2011.

[32] M. Filippini, "Short- and long-run time-of-use price elasticities in Swiss residential electricity demand," *Energy Policy*, vol. 39, no. 10, pp. 5811–5817, 2011.

[33] D. Livengood and R. Larson, "The energy box: Locally automated optimal control of residential electricity usage," *Service Sci.*, vol. 1, no. 1, pp. 1–16, 2009.

[34] D. Petersen, J. Steele, and J. Wilkerson, "WattBot: A residential electricity monitoring and feedback system," in *Proc. Extended Abstracts Human Factors Comput. Syst. (CHI)*, Boston, MA, USA, 2009, pp. 2847–2852.

[35] K. Wangpattarapong, S. Maneewan, N. Ketjoy, and W. Rakwichian, "The impacts of climatic and economic factors on residential electricity consumption of Bangkok Metropolis," *Energy Build.*, vol. 40, no. 8, pp. 1419–1425, 2008.

[36] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Proc. Workshop Data Min. Appl. Sustain. (SIGKDD)*, San Diego, CA, USA, 2011.

[37] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Secur. Privacy*, vol. 8, no. 1, pp. 11–20, Jan./Feb. 2010.

[38] S. Wicker and R. Thomas, "A privacy-aware architecture for demand response systems," in *Proc. 44th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Kauai, HI, USA, 2011, pp. 1–9.

[39] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol, CA, USA: O'Reilly, 2013.

[40] M. F. Porter, "An algorithm for suffix stripping," *Program Electron. Library Inf. Syst.*, vol. 14, no. 3, pp. 130–137, 1980.

[41] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Berkeley, CA, USA, 1999, pp. 42–49.

[42] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, 2004.

[43] M.-P. Kwan, "The uncertain geographic context problem," *Ann. Assoc. Amer. Geographers*, vol. 102, no. 5, pp. 958–968, 2012.

[44] D. W. S. Wong, "The modifiable areal unit problem (MAUP)," in *WorldMinds: Geographical Perspectives on 100 Problems*. Dordrecht, The Netherlands: Springer, 2004, pp. 571–575.

[45] G. Heinrich, "Parameter estimation for text analysis," Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2005. AQ9

[46] C. W. J. Granger, "Some recent development in a concept of causality," *J. Econometrics*, vol. 39, nos. 1–2, pp. 199–211, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0304407688900450

[47] H. H. Lean and R. Smyth, "Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia," *Energy*, vol. 35, no. 9, pp. 3640–3648, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544210002719

[48] R. J. Cabin and R. J. Mitchell, "To Bonferroni or not to Bonferroni: When and how are the questions," *Bull. Ecol. Soc. Ameri.*, vol. 81, no. 3, pp. 246–248, 2000.

[49] A. Aroonruengsawat and M. Auffhammer, *Impacts of Climate Change on Residential Electricity Consumption: Evidence From Billing Data*. Chicago, IL, USA: Univ. Chicago Press, 2011.

[50] A. Jacobson, A. D. Milman, and D. M. Kammen, "Letting the (energy) Gini out of the bottle: Lorenz curves of cumulative electricity consumption and Gini coefficients as metrics of energy distribution and equity," *Energy Pol.*, vol. 33, no. 14, pp. 1825–1832, 2005.

[51] S. Kishore and L. V. Snyder, "Control mechanisms for residential electricity demand in smartgrids," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Gaithersburg, MD, USA, 2010, pp. 443–448.

[52] K. P. Anderson, "Residential energy use: An econometric analysis," RAND, Santa Monica, CA, USA, Tech. Rep. R-1297-NSF, Oct. 1973. AQ10

[53] D. W. Caves and L. R. Christensen, "Econometric analysis of residential time-of-use electricity pricing experiments," *J. Econ.*, vol. 14, no. 3, pp. 287–306, 1980.

[54] J. K. Dobson and J. D. A. Griffin, "Conservation effect of immediate electricity cost feedback on residential consumption behavior," in *Proc. 7th ACEEE Summer Study Energy Efficiency Build.*, vol. 2. Pacific Grove, CA, USA, 1992, pp. 33–35.

[55] G. Lafrance and D. Perron, "Evolution of residential electricity demand by end-use in quebec 1979-1989: A conditional demand analysis," *Energy Stud. Rev.*, vol. 6, no. 2, pp. 164–173, 1994.

[56] I. Matsukawa, "The effects of information on residential demand for electricity," *Energy J.*, vol. 25, no. 1, pp. 1–17, 2004.

[57] M. Parti and C. Parti, "The total and appliance-specific conditional demand for electricity in the household sector," *Bell J. Econ.*, vol. 11, no. 1, pp. 309–321, 1980.

AQ11 AQ12

Todd Bodnar (M'XX) received the B.Sc. degree in computer science from the Pennsylvania State University, State College, PA, USA, in 2012, and the Ph.D. degree in biology in 2015.

His current research interests include machine learning and data mining on large datasets to measure sociological patterns.

Matthew L. Dering received the B.A. degree in psychology from Swarthmore College, Swarthmore, PA, USA, in 2007, and the M.S. degree in computer science from the Pennsylvania State University, State College, PA, USA, in 2014, where he is currently pursuing the Doctoral degree under the supervision of Dr. C. Tucker.

His research interests include computer vision, novel data sources, and video analysis, especially pertaining to sports.

Conrad Tucker (M'XX) received the B.S. degree in mechanical engineering from the Rose-Hulman Institute of Technology, Terre Haute, IN, USA, in 2004, and the M.S. degree in industrial engineering, the M.B.A. degree in business administration, and the Ph.D. degree in industrial engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA.

His current research interests include formalizing system design processes under the paradigm of knowledge discovery, optimization, data mining, informatics, applications in social media network mining of complex systems, design, and operation, product portfolio/family design, and sustainable system design optimization in the areas of energy, healthcare, consumer electronics, environment, and national security.

Kenneth M. Hopkinson (SM'XX) received the B.S. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1997, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, USA, in 2002 and 2004, respectively, all in computer science.

He is a Professor of Computer Science with the Air Force Institute of Technology, Wright-Patterson AFB, OH, USA. His current research interests include simulation, networking, and distributed systems.

AUTHOR QUERIES
AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

If you have not completed your electronic copyright form (ECF) and payment option please return to the Scholar One "Transfer Center." In the Transfer Center you will click on "Manuscripts with Decisions" link. You will see your article details and under the "Actions" column click "Transfer Copyright." From the ECF it will direct you to the payment portal to select your payment options and then return to ECF for copyright submission.

AQ1: Fig. 6 is cited before Fig. 4. Please check if Fig. 6 can be renumbered so that they are cited in sequential order.
AQ2: Please cite "Fig. 2" and "Table III" inside the text.
AQ3: Please provide expansion for the term "AIC."
AQ4: Please confirm that the location and publisher information for References [1] and [44] is correct as set.
AQ5: Please provide the location for Reference [10].
AQ6: Please confirm that the edits made to Reference [26] is correct as set.
AQ7: Please provide the page range for Reference [36].
AQ8: Please provide the issue number or month for Reference [42].
AQ9: Please provide the department name and technical report number for Reference [45].
AQ10: Please provide the department name for Reference [52].
AQ11: Please provide the membership years for the authors "T. Bodnar, C. Tucker, and K. M. Hopkinson."
AQ12: Please provide the organization name for the degrees attained by the author "T. Bodnar."

Using Large-Scale Social Media Networks as a Scalable Sensing System for Modeling Real-Time Energy Utilization Patterns

Todd Bodnar, *Member, IEEE*, Matthew L. Dering, Conrad Tucker, *Member, IEEE*, and Kenneth M. Hopkinson, *Senior Member, IEEE*

***Abstract*—The hypothesis of this paper is that topics, expressed through large-scale social media networks, approximate electricity utilization events (e.g., using high power consumption devices such as a dryer) with high accuracy. Traditionally, researchers have proposed the use of smart meters to model device-specific electricity utilization patterns. However, these techniques suffer from scalability and cost challenges. To mitigate these challenges, we propose a social media network-driven model that utilizes large-scale textual and geospatial data to approximate electricity utilization patterns, without the need for physical hardware systems (e.g., such as smart meters), hereby providing a readily scalable source of data. The methodology is validated by considering the problem of electricity use disaggregation, where energy consumption rates from a nine-month period in San Diego, coupled with 1.8 million tweets from the same location and time span, are utilized to automatically determine activities that require large or small amounts of electricity to accomplish. The system determines 200 topics on which to detect electricity-related events and finds 38 of these to be valid descriptors of energy utilization. In addition, a comparison with electricity consumption patterns published by domain experts in the energy sector shows that our methodology both reproduces the topics reported by experts, while discovering additional topics. Finally, the generalizability of our model is compared with a weather-based model, provided by the U.S. Department of Energy.**

***Index Terms*—Event detection, Granger causality, predictive models, social network services, unsupervised learning.**

Manuscript received January 11, 2016; accepted October 3, 2016. This work was supported by the Air Force Summer Faculty Fellows Program. This paper was recommended by Associate Editor Y. Wang.

T. Bodnar is with the Center for Infectious Disease Dynamics, Pennsylvania State University, State College, PA 16802 USA (e-mail: toddbodnar@gmail.com).

M. L. Dering is with the Department of Computer Science and Engineering, Pennsylvania State University, State College, PA 16802 USA (e-mail: mld284@cse.psu.edu).

C. Tucker is with the Department of Engineering Design, Pennsylvania State University, State College, PA 16802 USA, the Department of Industrial Engineering, Pennsylvania State University, State College, PA 16802 USA, and also with the Department of Computer Science and Engineering, Pennsylvania State University, State College, PA 16802 USA (e-mail: ctucker4@psu.edu).

K. M. Hopkinson is with the Department Electrical and Computer Engineering, Air Force Institute of Technology, Dayton, OH 45433 USA (e-mail: kenneth.hopkinson@afit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2618860

I. Introduction

Social media network models have the potential to serve as dynamic, ubiquitous sensing systems that serve as an approximation of physical sensors with the added benefits of: 1) being scalable; 2) publicly available; and 3) having lower setup and maintenance cost, compared to certain physical sensors (e.g., smart meters or smart plugs). Each day, social media services such as Twitter, Facebook, and Google, process anywhere between 12 terabytes (10^{12}) [1] to 20 petabytes (10^{15}) [2] of data, making them suitable for large-scale data mining and knowledge discovery. The ability of individuals within a social media network to: 1) detect a phenomenon; 2) observe and interpret a phenomenon; and 3) report the impact of the phenomenon back to the social media network in a timely and efficient manner, highlights the potential for social media networks to be perceived as large-scale sensor networks. However, as with many large-scale sensor systems, the fundamental challenge is separating signal from noise. The conventional wisdom has been that in order to accurately understand a complex phenomenon (e.g., energy utilization patterns), complex sensors are required (e.g., smart meters) to sense, collect data, and make inferences in real time. This paper aims to challenge these conventional paradigms of social media networks and physical sensor systems by demonstrating the viability of social media networks to be used as dynamic, ubiquitous sensing systems that provide comparable level of information and knowledge, to physical sensor systems setup to achieve similar objectives.

In this paper, we propose a system that automatically generates and tests relationships between topics on social media network and electricity usage pattern. These topics are then used to predict future electricity use or test Granger causal links between the topics and the usage. This Granger causality is used to validate these links. We consider a case study where our methods are applied to energy use disaggregation using social media network data. That is, can our system discover interesting relations in social media networks that trend with electricity consumption rates? We then compare the topics that our system detects to be valid against actual topics chosen by an expert in the energy domain or against keywords mined directly from the dataset. We find that, in addition to other topics, our system replicates the topics chosen by an expert. Furthermore, a direct comparison to keyword analysis results in up to a 16.7% improvement in detected correlations (as described in Section V-B). Finally, a comparison with a weather-based simulation of homes in cities is considered.

In this paper, we provide an implementation, quantitative evaluation, and analysis of this mapping. In Section II, previous work on social media network analysis, topic modeling, and electricity use disaggregation is discussed. In Section III, a formal implementation of this mapping system is provided. In Section IV, a case study is presented where $y =$ *electricity consumption rates,* and $\mathbf{X}$ is statistically derived social media network data. In Section V, this method of hypothesis generation is compared against expert-based and machine learning-based hypothesis generation. In Section VI, we test our model's capability to predict future electricity usage. In Section VII, we conclude.

II. Previous Work

A. Mining Social Media Networks

Social media networks are emerging as the next frontier for novel information discovery. Previous work has shown applications toward measuring weather patterns [3], diagnosing illness [4], tracking earthquakes [5], providing user recommendations [6], exploring plans of action in crises [7], detecting security risks [8], and describing obesity patterns [9]. Part of social media network's advantage is the relatively openness and ease of collection of data, which, unlike traditional websites, are created by a larger population of users whose demographics are more representative of the general population [10].

One way that social media network data can be represented is as a set of sensors, where each user is a noisy sensor [4], [5]. That is, instead of reporting numerical data like traditional sensors do, social media network users report textual data which must be preprocessed before statistical methods can be applied. Simple keyword analysis—a mainstay of modern text analysis—can be problematic when applied to big datasets. For example, Google Flu Trends' system of applying text analysis to search queries has been shown to over estimate ground truth influenza rates [11], [12]. In this paper, we employ topic modeling to avoid the worst case scenario of an exhaustive search of keyword-phenomena relations.

B. Topic Modeling

Topic modeling is a way to algorithmically derive topics from unstructured documents of text. Modern work has been focused on latent Dirichlet allocation (LDA) and its derivatives [13]–[15]. LDA works by determining clusters of words in a document to determine "topics" through a Bayesian process. These topics can be represented by the words that, statistically, best describe the cluster. It has been shown that LDA can be used to detect topics in datasets such as Wikipedia articles [16], [17], scientific literature [18], spam classification [19], news analysis [20], and tweeting behavior [9], [21]. In this paper, we demonstrate that the set of topics generated by topic modeling algorithms are indeed statistically valid approximations of events. We further show that by mining these event-phenomena patterns, researchers can discover events strongly related to phenomena of interest.

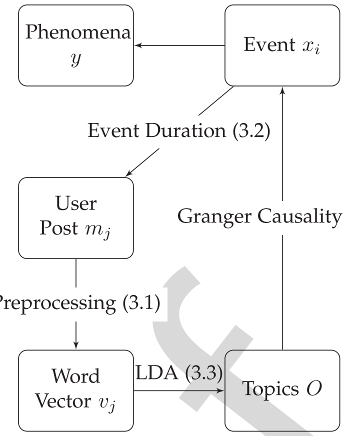

Fig. 1. High-level description of our system to transform a social media network stream into hypotheses about a real world event.

C. Knowledge Management in Energy Systems

Smart grids use communication to facilitate context awareness and cooperation across much wider areas than previous power grid system [22]. Among the initiatives to introduce smart grids are the advance metering infrastructure [23], [24] for metering in the distribution system, demand pricing, IEC 61 850 substation automation [25], the wide area management system [25], [26] for wide-area PMU measurement, and the North American synchrophasor initiative [27] that uses wide-area utility communication. Smart grid operations rely on periodic collection of data through sensors followed by processing the data.

The technology provided by a smart grid is valuable for reducing or predicting large spikes in electricity utilization. For example, by coordinating households to not perform high power usage activities concurrently. However, the smart grid has not yet been widely implemented. This paper has focused on methods to study nonsmart grid data to study either high-level usage patterns, such as total energy consumption in a city, low-level usage patterns to measure device level energy consumption, or placement of systems based on simulation [28]. It would be difficult to generalize high-level measurements to work at a finer grain because real-time electric consumption sensors are typically deployed on a station or node level. Thus analysis is limited to events that impact a large area, such as the temperature or time of day [29]–[35]. Low-level, device-based measurements have been proposed as a method to disaggregate high-level power consumption patterns [36]–[38]. These sensor networks have the advantage of providing device-level information and bypassing the need to rely on a power company for data. However, these are expensive to implement and require installation of hardware in the study participant's house, limiting the amount of data that can be collected.

To demonstrate the practicality of our system in real life applications, we consider applying our system of automated event detection to provide a novel system of energy usage disaggregation which can take high-level, publicly available power consumption records and generate valid hypotheses about behaviors that affect this consumption. For a graphical description of our methods (see Fig. 1). First, we clean textual social media network streams. Then we use LDA on the cleaned text to detect topics. These topics are then used as the basis for hypotheses about a real-world event. These topics are then tested for statistical significance. Validated hypotheses are then reported.

III. Social Media Network Electricity Utilization Methodology

In this section, we propose using large-scale social media network data as method of tracking a subset of events that are relevant to the social media network users, $\mathbf{X}$. That is, exposure to a particular event $x_i \in \mathbf{X}$ may induce a user to post a message m at time j, m_j on a social media network. Here, we assume m_j to be text-based. That is, it can be represented with a word vector v_j, derived from the raw message m_j. While it is easy for a user to map $x_i \rightarrow m_j$ (for example, “I need to do my laundry”), it may be hard to reverse this mapping, at least in a machine processable manner. Since our goal is to generate these x_i to test against phenomena, in this case: electricity usage, we must approach this mapping in an indirect fashion. Thus, we develop topic models from these word vectors where we assume a topic o is an approximation to event x_i for some i. Later, we provide an empirically tested and validated analysis of this assumption (see Section V). This allows us to map $m_j \rightarrow v_j \rightarrow o \rightarrow x_i$, effectively reversing the mapping of $x_i \rightarrow m_j$ in an unsupervised manner. Thus, we are able to formulate and validate statements of the form “x_i is related to phenomenon y” without prior knowledge about x_i.

A. Cleaning Raw Social Media Network Data

Social media network data are commonly described as extremely noisy [3], [5], [39], requiring intensive cleaning of the social media network stream as a necessary first step. We do this by converting a string of characters into a list of n-grams—pairs of up to n contiguous words (see Algorithm 1). The n-grams are determined by tokenizing the string on all nonalphabetical characters. Since capitalization can be erratic in social media networks, the n-grams are then converted to lowercase. As the objective of this step is to derive topics instead of keywords, we stem each of these words using porter stemming [40]. This maps words with similar stems but with different suffixes to the same keyword. For example, “accept,” “accepting,” and “acceptance” are all mapped to the same keyword, accept.

This list of n-grams is expected to follow a long-tail distribution [41], resulting in the likelihood that some are too common or too rare to be valuable in the analysis. Common words such as “the,” “is,” and “and” give little or no information about the text and could overshadow other, more descriptive, words that do not occur as frequently [13], [17], [42]. Thus common words, as defined by Lewis *et al.*’s [42] *stop list*, are removed from the list of n-grams. On the other hand, if a word is too rare, it may not occur enough for any inferences about it to be generalizable. Since the distribution of n-grams has a long-tail, most words will be too rare. Thus there is the potential of these very-rare n-grams to lower our ability to generate inferences about *any* n-grams [4], [17], [41]. This problem is addressed by removing any n-grams that occur less than c times [4], [17], [20], [41], [42]. However, previous work tends to be somewhat vague on how to determine c, often incorporating expert knowledge to determine c. Here, we determine c algorithmically.

To determine c, first begin with the distribution of n-gram counts. That is, f_m is the number of n-grams that occur exactly m times each in the dataset. We then iteratively test each value for $c > 0$ until we find the minimum value for c such that

$$\frac{f_c}{\sum_{m=c+1}^{\infty} f_m} < \delta_{\min} \tag{1}$$

where $\delta_{\min}$ is a user defined stopping threshold. Thus, we define rare words as words that occur less than δ times and remove them—a necessary step for preprocessing for LDA [17].

Note that we specifically do not remove keywords related to URLs as they may provide additional information about the user’s activity. For example, tweets containing a link copied from a Web browser are likely to include “http” which may be less common on mobile users. Alternatively, links with “4sq” (reduced to “sq” when numerics are removed) are sent through four square’s—a popular location check-in service—mobile application, informing us that the user is more likely visiting a location outside of his or her house.

B. Pairing Real World and Social Media Network Data

Social media network data can be updated on a millisecond level; however, it is rare for real-world events to be reported at such a temporal resolution. Additionally, it is unlikely that a single social media network message contains significant, relevant information about the real-world event we want to study, or if it does, they are exceedingly rare. We address this discrepancy by normalizing the social media network data to the real-world data’s time scale. That is, we define a document

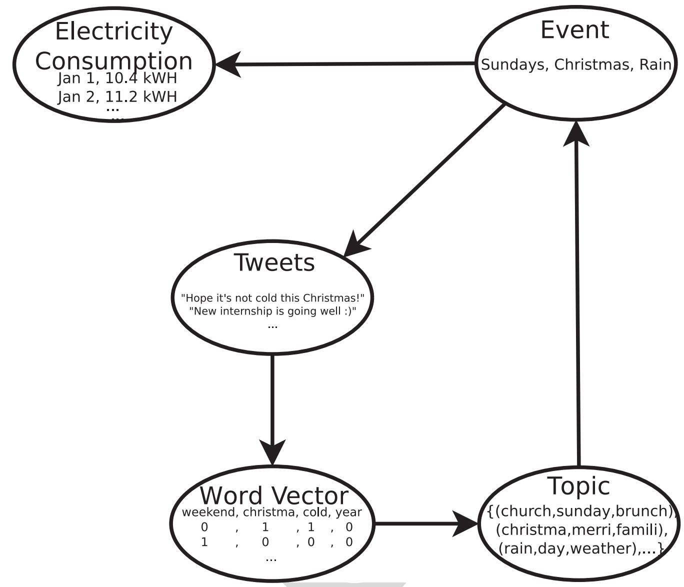

Fig. 2. Implementation of our theoretical model (see Fig. 1) for our case study.

Algorithm 1: Preprocessing Steps for Social Media Network Data

Data: Time tagged Messages $\mathbf{M}$
Result: A set of aggregated and processed messages $\mathbf{D}$
d_q = document of keywords at time q;
$count_{word}$ = frequency of "word" in all documents;
$\mathbf{W}$ = set of all known stemmed words;
for $\mathbf{m_j} \in \mathbf{M}$ **do**
  Break $\mathbf{m_j}$ into substrings on non-alphabetical characters ^$[a - zA - Z]$;
  $\mathbf{j}$ = hour $\mathbf{m_j}$ was posted;
  for *non-empty Substring* $\mathbf{S}$ *in* $\mathbf{m_j}$ **do**
    convert $\mathbf{S}$ to lowercase;
    stem $\mathbf{S}$ using porter stemmer;
    add $\mathbf{S}$ to $\mathbf{W}$;
    push $\mathbf{S}$ onto $d_{\mathbf{j}}$;
    $count_{\mathbf{S}}$ ++;
  end
end
for *word* $\mathbf{S}$ *in* $\mathbf{W}$ **do**
  if $count_{\mathbf{S}} < \delta_{min}$ **then**
    Remove $\mathbf{S}$ from each d_q;
    Remove $\mathbf{S}$ from $\mathbf{W}$;
  end
end

Algorithm 2: LDA Algorithm in the Context of the Proposed Social Media Network Model

Data: set of **Documents** D, topics $\mathbf{O}$
Result: a $|\mathbf{W}| \times |\mathbf{O}|$ matrix
for Document $d \in D$ **do**
  for *Word* $w \in d$ **do**
    w_{topic} = Random topic $\in \{0, \ldots, |\mathbf{O}|\}$;
  end
end
for *Step in* $\{1, \ldots,$ *stop point*$\}$ **do**
  for Document $d \in D$ **do**
    for *word* $w \in d$ **do**
      for *topic* $o \in \{0, \ldots, |\mathbf{O}|\}$ **do**
        $P(o|d) = \frac{|w \in d \text{ where } w_{topic} = o|}{|w \in d|}$;
        $P(w|o) = \frac{|w \in D \text{ where } w_{topic} = o|}{|w \in D|}$;
      end
      Assign w_{topic} based on $P(w|o) \times P(o|d)$.
    end
  end
end

d_j to be the aggregation of all processed social media network messages v_j (as derived from m_j) that occur in during the timespan between the qth real-world event x_q and the next event, x_{q+1}. More formally

$$d_j = \{v_j | \text{time}(x_q) \leq \text{time}(m_j) < \text{time}(x_{q+1})\} \tag{2}$$

where v_j is the n-gram representation of message m_j and time(e) is the time when e occurs. For example, if one is looking at temperature data that is reported on an hourly basis, a document would be all posts that occur within that hour. Algorithm 1 outlines how these messages are processed into word vectors, and subsequently aggregated into a document. It would be unreasonable to assume that a user posts a message *exactly* when the event happens. Instead, it is likely that the user posts about an event sometime *before*, *during*, or *after* the time that the event occurs. This issue is partially addressed when the data is aggregated, because all message after an event, but before the next, will be combined, regardless of lag between event and message.

Additionally, data can be paired based on geospatial information, such as which zip code the message occurred in. This is dependent on the dataset describing the phenomena y and the social media network messages $m_j \in \mathbf{M}$ both containing comparable location data. Caution should be advised if arbitrary spatial units are defined: the "modifiable areal unit problem" can bias results from geospatial aggregation and remains an open problem [43], [44].

C. Generating Topic Models

A given set of documents defined by the aggregation described above can be used to generate topics through LDA. We use Gibbs sampling [17], [18] implemented by JGibbLDA [17] to perform this analysis. LDA determines the probability of a document being about a topic given that it contains a set of n-grams [13], [17], [18]. To do this, LDA first generates clusters of words based on co-occurrence in the documents. That is, the probability of a word w occurring given that a document is in topic o_w. To represent these topics in a human readable form (for example, in Tables I and II), we present the set of words that have the highest probability of occurring within the topic. In other words, the topics can be expressed as a $|\mathbf{W}| \times |\mathbf{O}|$ matrix, where $\mathbf{W}$ is the vocabulary found in Section III-A and $\mathbf{O}$ are the topics generated by the LDA model such that $\mathbf{o} \in \mathbf{O}$. Each entry in this matrix corresponds to the probability of that word belonging to that topic. LDA works according to Algorithm 2. Note that the stop point is selected as 2000, the default of JGibbLDA as proposed by Heinrich [45]. This algorithm uses as input each of the aggregated Documents from Algorithm 1 to generate O topics.

The probabilities contained in this matrix can be reversed using Bayes' theorem to determine the probability that a document is in topic o given that it contains a set of keywords. Since each document has a related time component, we can say that the probability of a document being in o varies over time. By considering the likelihood of all topics over all documents, we can observe the changing interests of the population of users over time. Each of these topics are the basis of a question: "Question: Is the ith event x_i (as inferred from topic o) related to real world phenomena y?"

D. Determining Event-Phenomena Causality

In Section III-C, we outlined the method to generated topics—which we later show in Section IV-C to be statistically

Algorithm 3: Mapping Topics to Effects

```
Data: Documents D and Topics O from Algorithms 1
      and 2
Result: Granger Causal Topics
for document d ∈ D do
    for topic o ∈ O do
        TS_{o,d} = rate of o in d
    end
end
for o ∈ O do
    Significance = Granger(TS_o,PowerUsage);
    if Significance then
        Print o;
    end
end
```

Algorithm 4: Computational Complexity of This Methodology

```
input : Social Media Posts
output: Predictions
Social Media Posts arrive: O(1);
Preprocessing: O(m) where m = number of posts;
topics ← Generate Topics (LDA): O(Nm^2) (see alg 2);
CausalTopics ← Granger(topics) O(Len(topics)) ;
```

valid approximations of events—from social media network and determined the frequency of each topic at a given time. Next, we explore the patterns of each of these events over time. That is, combining all frequencies of an event over time results in a time series to be compared to the real world phenomena. Some topics, such as *Christmas*, *hating Mondays*, or *having lunch* will display cyclical patterns while other events, such as ones about a *hurricane* or a *concert*, may be one-time, anomalous events.

The event's time series can be compared to the document time series related to the real-world phenomena through cross-correlation (see Algorithm 3). That is, by matching events frequencies and real world phenomena by their time, can we find any relations between the two variables? This is defined by the Pearson's rank correlation where each point is a pairing of event frequencies and real world phenomena. The system does not filter by positive or negative correlation: a strong negative relationship between an event and a real world event can be just as interesting as a positive one. While these correlations may be strong, they do not necessarily imply a causal link.

While we do consider a correlative analysis between automatically detected events and electricity consumption, there is also an interest in determining which—if any—of the behaviors have a causal relationship on the electricity rates. Detecting strong causality through an uncontrolled, observational study without an external model of the system is impossible. Hence, we focus on detecting Granger causality [46], [47], a less stringent form of causal testing. Simply put, "correlation does not imply causality" because there may be a third phenomena that influences both, or if there is a causal relation between the two phenomenas, it is impossible to tell which one causes the other without external information. Granger causality addresses the second issue by employing lagged data. This aids in establishing a causal relationship by testing not only the synchronous variables, but measuring if the lagged data aids in the explanatory power of the model. That is, can information about phenomena y at time t (y_t) be inferred by a behavior x at time $t-t'$, for some positive value of t'? If it can, then we at least know which direction causality is flowing. To control for auto-correlative effects, the standard model compares an auto-correlation model of the predicted phenomena y

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_{(\text{lag}_{\max})} y_{(t-\text{lag}_{\max})} \quad (3)$$

where $\text{lag}_{\max}$ is the maximum lag considered in the model, determined by maximum likelihood estimation. We then add the lagged components from an event's trend x_i to the formula

$$\begin{aligned} y_t = {} & \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_{(\text{lag}_{\max})} y_{(t-\text{lag}_{\max})} \\ & + \beta_{(2+\text{lag}_{\max})} x_{i,t-1} + \ldots + \beta_{(2*\text{lag}_{\max})} x_{i,(t-\text{lag}_{\max})}. \end{aligned} \quad (4)$$

The predictive power of these two models is compared by performing a t-test on the errors between the two models. If we find that (4) performs better than (3), then it is because knowledge about this second event informs us about the future state of the target phenomena. While this is still not a test for true causality, Granger [46] have argued that it is a step in that direction. Note that Granger causality does not control for a third phenomena, which influences both the ith x in question, x_i, and y, other than guaranteeing that it occurs at some point before y. Indeed, in our case, we assume that a behavior influences both x_i—tweeting about the risk factor—and y—later power consumption due to the behavior. This method of dual time series analysis has two benefits: it quantifies how long of a lag is meaningful, and determines which sampled topics are significant.

This Granger causal test allows us to quantify the causal relationship between a phenomenon (a change in power usage) and an event (as represented by one or more social media topics). This causality measurement is the primary method of establishing causality implemented in this methodology. Social media posts can be processed into topics ahead of time, and these topics can be detected within new posts in linear time. This also allows these causal relationships to be updated in an online fashion. If the performance of the predictive nature of these causal relationships degrades, a new sample can be drawn and recalculated (see Algorithm 4). This allows us to adapt and use new data instead of relying solely on old data.

E. Validating Event-Phenomena Relationships

At this point, we have generated relationships of the form: "Topic o is related to a real world phenomena y with correlation r_o." However, if the coefficient of determination, r_o^2, is small, then any trends detected may not be statistically significant. Thus, we calculate the p-value for each regression. Since the system may test hundreds or thousands of regressions,

Algorithm 5: Topics to Predictions

Data: Significant Time Series S, **PowerUsage** Data
Result: Measurement of Predictive value of Social Media Network data
Let $\mathbf{PowerUsage}_h = \mathbf{PowerUsage}$ data lagged by h hours;
$\mathbf{S}_{a,b} = a^{th}$ significant Time Series lagged by b hours;
Build model $f(\mathbf{S}_{a,b}, \mathbf{PowerUsage}_b) = \mathbf{PowerUsage}$;
Evaluate f on data from subsequent time period;

the traditionally chosen cut off $\alpha = 0.05$ must be corrected. That is, if 100 tests are conducted on randomly generated data, it is likely that five will be reported as false positives. Bonferroni correction [48] was chosen because it does not depend on normal distribution or independence assumptions. Bonferroni correction defines the corrected cut off as $\alpha' = \alpha/n$, where n is the total number of hypotheses tested. This method of correction is more conservative than others, giving more assurance that any hypotheses that do pass the test are valid.

By implementing our system, events can be inferred from social media network data which can inform researchers about real world phenomena, as we will show in Sections IV–V. Finally, we evaluate the predictive value of this methodology as outlined in Algorithm 5 in Section VI.

IV. Case Study

In this section, we demonstrate the feasibility of our system on Twitter data, in order to determine whether topics can help explain a real world energy utilization (see Fig. 1). Specifically, we consider electricity consumption from single-family households in San Diego County from March 3, 2011 to December 31, 2011 and 1.8 million tweets from the same timespan that originated from San Diego County. That is, $\mathbf{M}$ ={Tweets in San Diego between March 3, 2011 and December 31, 2011} and y = *electricity consumption rates in kilowatts.*

A. Description of Datasets

Electricity consumption data was provided by the San Diego County Gas and Electric Company which supplies power to residents of the San Diego County in southern California. Data was provided on a daily basis for the year 2011 and represents a typical, single-family, residence.[1] Power usage data was discarded before the initial collection of Twitter data on March 3. Since power usage has both a daily cycle and longer-term dynamics (see Fig. 3), we consider both hourly and daily aggregation of the data.

Twitter data was collected between March 3, 2011 and December 31, 2011 through the Twitter API by searching for all tweets with high-resolution geospatial data. Additionally, tweets are filtered to be located within San Diego County as

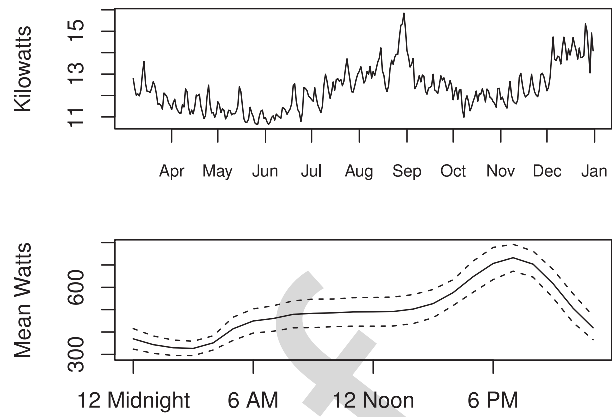

Fig. 3. Mean daily and hourly rates of power consumption for San Diego residents. Dashed lines in hourly graph indicate one standard deviation.

defined by the 2011 TIGER shape file[2] to match the spatial boundaries of the power data. A total of 1 813 689 tweets matched this criteria. The raw *jsons* returned by the Twitter API were then processed through The Open Twitter Parser[3] and stored in an MySQL database for further processing.

B. n-Gram Selection

Next, the Twitter data was cleaned through tokenization, stemming, case-normalization and *stop word* removing, as described in Section III-A. In this case study, we only consider unigrams (n-grams where $n = 1$) for analysis. Only unigrams are considered for several reasons. A higher dimensionality would cause the number of correlations to be calculated to explode. Furthermore, most implementations of LDA only consider unigrams, as topics must be latent relationships between these unigrams. Finally, given that the dataset of length constrained social media posts is studied, it is fairly common for users to discard words, which would severely limit the usefulness of n-grams for $n > 2$. A total of 794 917 unigrams were detected. We set $\delta_{\min}$ to one percent and determine that the optimal cut c to be 102, removing any unigrams that occurred less than 102 times in the dataset (see Fig. 6), AQ1 thus we define $\delta_{\min}$ and use this to calculate c, which allows our approach to scale to datasize. This automated selection of c generates comparable results to other papers [16]–[18] that use domain knowledge to choose their cut off, while still allowing for more or less frequencies depending on datasize. This also helps if new samples need to be drawn and tested.

C. Knowledge Discovery of Statistically Relevant Social Media Topics

We now aim to show that these topics are statistically related to the real world events that they describe, as our assumptions in Section III require. As a null hypothesis, we consider that individuals are free to discuss any topic at any time. That is, the probability of a topic being discussed, $P(o)$ does *not* depend on the time. Instead, if our original assumption is correct, then $P(o\|x_i) > P(o)$ for some $o \in O$ and $x_i \in \mathbf{X}$. Hence, a

[1] http://www.sdge.com/sites/default/files/documents/Coastal_Single_Family_Jan_1_2011_to_Jan_1_2012.xml

[2] http://www.census.gov/geo/www/tiger/tgrshp2011/tgrshp2011.html

[3] https://github.com/ToddBodnar/Twitter-Parser

TABLE I
WORDS THAT BEST DESCRIBE THE 20 DAILY TOPICS FROM TWITTER THAT OUR SYSTEM DETERMINED TO BE ABOUT POWER USAGE AND THE CORRELATION BETWEEN THE TOPIC AND POWER USAGE. NOTE THAT THE TOPICS HAVE BEEN SORTED BY CORRELATION COEFFICIENT

r	Most likely words in the topic
-0.519	**job** http ly bit ca sandiego tinyurl **getalljob** www **tweetmyjob** lt **manag** electron soni **service** carlsbad gt ...
-0.480	sq http **instagr** la gowal ly bit job san diego tinyurl **twitpic** es lt **beach** sandiego **day great foursquar** www ...
-0.344	**jobcircl cybercod job** ca **engin develop hire softwar** mesa **sale** www **senior** la **design manag** net voic **game web** ...
-0.335	work gt dr check street **offic** fit show diego **hour** center **facebook starbuck art** airport media mesa **lunch** busi ...
-0.301	rt **coupon summer** spag june caseyanthoni es sandiego **lockerz** souther em poway doi gov earthquak ...
-0.282	rt lmao **june** tinyurl spag **marathon** jonez **job getalljob** upling samoan rock roll heat damsel untp final show ...
-0.281	**weekend** spag coupon **memori** back cri **sad hangov disapoint** kck **justinbeib** sandiego es oprah lockerz support ...
-0.247	wednesday **fat** thrusday **muscl** free bit **weight** ur loss hump **diet wine** market fan friday set eleddieg hot ...
0.201	glass **sun** auto **sprinkler** rek rt **repair** xd **replac** tcot pancak commanderlov pae coupon del word mar ...
0.211	**real pretend** jlh thereal point don itsatumblrth year iamlaceychabert laceyoffici **handbag design manufactur** ...
0.225	jlh **frenchfan** victoria **witter clalovehewitt** lol alexandria thereal don rt tweet es **coupon** ya bcuz camill game ...
0.225	**christma cold** year dat jus sir final ass wyd bro si lo nba yea smh man dnt crystal twitter laker victoria ...
0.238	yummi day **sexi orgasm** good morn email hotmail saturn great **love beauti** school video lick class cum pretti ...
0.254	**christma merri famili eve xmas happi holiday santa** present lt **gift** year stephazilla laker **church navidad** hous ...
0.267	de eu pra um na da se vou mas uma mai meu tem vai em ver happi dia por ele minha person didn beauti ...
0.297	http san diego love **good time** day don make today back la job ca people haha lol **home** feel ll wait great ...
0.369	http **vista shop plaza valley** chula center **mall fashion buy** bonita peopl mission pkwi **store** home break work...
0.410	lt gt **lol fuck shit** job haha ca dr **ass bitch** don sandiego nigga **hate** girl feel love sleep drink ave **damn** ...
0.418	rek beauti **window hurrican** coupon iren xd vma video arhhhhjay lt omg **storm hot wind** gaga kk issu **humid** ...
0.448	lol **christma** final **holiday** dannyboyo partic travcb **home** laker deniseexclus **xmas** studi **happi** andruee ll ...

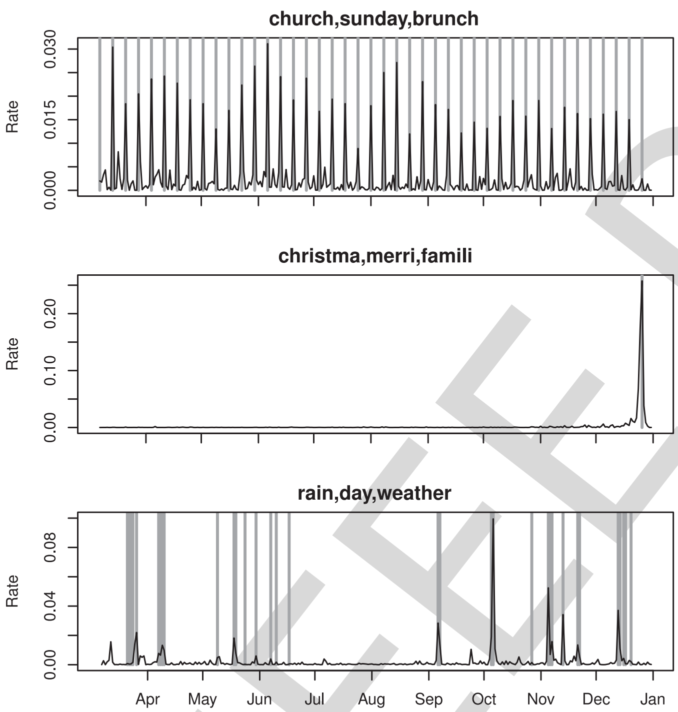

Fig. 4. Temporal patterns for three select topics in black. Chart titles indicate the three most representative words for each topic. Red lines indicate days that are Sundays, Christmas day, and days when it rained, respectively.

topic is being induced by a real world event. To verify this, we would need ground truth data for x_i, which is not necessarily possible to obtain for all possible events i. However, some topics lend themselves to easy validation.

Here, we consider three topics that appear to represent Sundays, Christmas, and rain (see Fig. 4). Sundays were chosen as the first topic for analysis because it was expected to follow clearly defined temporal patterns. Additionally, Sundays are more discrete than Christmas (i.e., are Christmas Eve and Christmas separate events?) or rain (i.e., how much precipitation is necessary for it to be considered raining?). Christmas was chosen as a topic because it is representative of events that occur only once in our dataset, but with a well defined event time. Indeed, Christmas may be the biggest topic detected, with 25.8% of the Twitter data being about Christmas on Christmas, December 25, and 17.5% on Christmas Eve, December 24. Finally, we considered rain because it lacks the periodicity of the other two topics. Note that the rate of precipitation does not have a strong relationship to spikes in the rain topic, so we discretized weather into days without rain and days with rain, as defined by weather underground.[4]

We can thus calculate the relevant probabilities (see Table I). This means that 80 topics whose correlations are too low are not present in this table. For example, with the topic sunday: $P(o_{\text{sunday}}) = 0.00350$ (as determined by LDA) and $P(o_{\text{sunday}}|x_{\text{sunday}}) = (o_{\text{sunday}}\&x_{\text{sunday}}/x_{\text{sunday}}) = 0.0182$. For completeness, we can use Bayes theorem to determine the probability that it is Sunday given that the topic is about Sunday

$$P\left(x_{\text{sunday}}|o_{\text{sunday}}\right) = \frac{P\left(o_{\text{sunday}}|x_{\text{sunday}}\right)p\left(x_{\text{sunday}}\right)}{P\left(o_{\text{sunday}}\right)}$$

$$= \frac{0.0182 * 0.15}{0.00350} = 0.728. \quad (5)$$

Since we know what days we sampled from, we know that $P(x_{\text{sunday}}) = (x_{\text{sunday}}/x_{\text{all}}) = 0.14$, which is close to the general occurrences of Sundays, (one out of seven days each week $\approx$0.1429). We find that $p(\text{Event}|\text{Topic})$ is significantly higher than the baseline $P(\text{Event})$, giving evidence toward these automatically generated topics, $o \in O$ having some relation to real world events $x_i \in \mathbf{X}$.

[4]http://www.wunderground.com/history/airport/KSAN/2011/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2011&req_city=NA&req_state=NA&req_statename=NA

AQ2

TABLE II
WORDS THAT ARE MOST ASSOCIATED WITH THE 38 HOURLY TOPICS FROM TWITTER THAT DESCRIBE EVENTS THAT ARE FOUND TO GRANGER CAUSE CHANGES IN POWER USAGE

r	Most likely words in the topic
-0.432	**job** http sandiego electron ca soni sd **sonyjob tweetmyjob engin** snei **softwar director** test alskks **develop administr** . . .
-0.321	lol shit yo lmao **work** man ass good nigga dat fuck smh tat ya feel dnt de jus bitch bro **sleep** je wit home est sir tha yea . . .
-0.145	**watch movi show** love time lol good great ll fun **tv** back night yeah make year peopl awesom **episod youtub** wait tweet . . .
-0.120	job http ca sandiego kaiser **nurs tweetmyjob healthcar** permanent san diego rn kindr **hospit** ii amn **kinderedjob** account . . .
-0.106	http esriuc love lol **harri** ddlovato rt **potter** time fstk googl good day kooldudestillo pride **watch** diego rhenderson demi girl . . .
-0.086	http rt **shop** lol great san www diego **ad sale** love lmao watch june item daili don day back mile summer **inventori** time good . . .
-0.079	http **california southern earthquak** gov km usg doi june **depth** usa diego gmt hour ca mi ll good time hand join monday. . .
-0.070	http el la ma love day al ya ben de ne play ana ve wait ha lol shit da good hey ni bi man check home ik en ba wo in tweet . . .
-0.057	**lol haha love** stephazilla **good** lt **hahaha** time watch don yeah fuck night feel back thing shit girl life wait tomorrow . . .
-0.057	**victoria witter** alexandria **teamjlh** stillo http **clalovehewitt** lol stellix don back good yeah tweet **beutyqueen** gonna . . .
-0.041	http del diego san mar la **fair beach blvd counti** day school jimmi ca **camino** de pic coronado vall durant time . . .
-0.035	http **japan** www greeney san **fukushima** good rt time **nuclear** ur win day **tsunami** great plixi ipad watch diego bit . . .
-0.029	http **plaza** diego san el citi bonita horton **shop** nation **westfield** hlbd cajon ave la parkway camino time de **mall** dr **buy** . . .
-0.027	**charger** http **game** diego san qualcomm **footbal statium win play raider** good team **watch** fan **nfl** time river **tebow** rt sunday . . .
-0.021	http diego san **coronado beach hotel** mission **bay** st pic pine del torrey **resort** la ave time spa park foursquar blvd vista . . .
-0.011	**work** make today rt **offic** ll **busi** free deal market don great health week **stori** peopl year **school** citi pay list design site news . . .
-0.003	jlh thereal frenchfan love real **jennif clalovehewitt** verifi lol **hewitt** http lt fake account don tweet back good day camill make . . .
-0.003	http **day lol love** diego back don time san ca good final ll class **cold** make **break** fuck work **night** week hate haha xoxo uni . . .
0.004	np love **song** shit make fuck don back real peopl **good music** lil man girl show **listen** thing yeah **play** damn haha rt . . .
0.006	**job getalljob** ca tinyurl sandiego http engin **edit manag telecommut concierg clinic assic** sleep hotel **remot develop web** hour . . .
0.018	na ko sa hahaha haha mo ako ng ang ka lang pa naman time eh day lol ba nga good si ni oo hehe hahahaha tweet . . .
0.023	**sleep night bed goodnight tomorrow** fuck good **dream** time **tonight** wake home **asleep** hour feel love drunk sweet happi . . .
0.027	job http ca general ga poway asi atom sys **aeronaut account sandiego tweetmyjob manufactur analysis** ii iii bit **financi control** . . .
0.029	te si de la ya tu mi el esta yo como en por lo se es para mas mero hola bien con bueno muy dia una todo ke los saludo pue . . .
0.048	de http la enl en los mexico se al del lol es para funal fuck love lt work por con su son home man tv mas twitter ha una las . . .
0.077	http juli happi day don **cassey caseyanthoni** good miss san make **sagesummit** time firework ll beach life peopl bit . . .
0.081	**game rt laker** lol **win** http heat **play watch team nba** fan love final good fuck lt day **season** bull **player** ve tonight time **kobe** . . .
0.089	http **life ratio live** tune proof net diego **fit good** back time tomorrow html **work** love guy night em cujo st lol miss watch . . .
0.100	rt http time ya love lol day teamfollowback di famili yg make **cricket** ll **wireless** ur gt good haha followback yo cool . . .
0.143	http **obama dead** diego **bin** san good **war** love news time **presid laden** rt **kill** day **cnn osama** de stop vote happi . . .
0.151	http san diego lunch st ave dr pic **cafe** blvd **grill food mexican** day **burger** today mayor **taco** work foursquar offic . . .
0.172	**iphon appl steve job** app rt live http today don rip twitter wait work feel **phone die** tattoo life love **ipad** world yeah io tweet . . .
0.175	**sq instagr** gowal la ly bit **twitpic foursquar** untp **mayor beach** trendsmap street lockerz tinyurl www btw picplz year . . .
0.184	http **today morn breakfast** san **church** diego **day cafe coffe** night **sunday good starbuck** park st hour mayor pic . . .
0.195	http **morn** san diego day **good today starbuck school earli work coffe** st fit oceansid carlsbad happi blvd wake mesa . . .
0.210	http san diego st park ave fan street south experi **hotel** tomorrow **intern** year ca gaslamp fun ll **rememb** market space . . .
0.226	http san diego st washington ave **chicago** btwn street el **game pizzeria pizza** map blvd fort good **cajon** lefti . . .
0.639	san diego http **airport** intern dr **termin harbor back** work **home** hour **flight** fit earli head line great **gate miss** begin . . .

Additionally, some events will show cyclical, daily patterns (see Fig. 5). If the target phenomena also shows similar patterns, these hourly events may further help to describe the phenomena.

D. *Event-Electricity Usage Relationships Detected*

These automatically determined topics were found to correlate with daily power consumption rates with $-0.519 < r_i < 0.448$ (see Table I). The topic that correlated most negatively with power consumption included unigrams such as "job," "getalljob," and "tweetmyjob." This leads to the first steps of a domain expert investigating that people use less energy at their residence on days when they are at work than days when they are not working. The topic that correlated most positively with power consumption included Levins stemmed unigrams such as "christma," "holiday," and "home," hinting that people consume more electricity around Christmas time. Similarly, the topics that were determined to Granger cause changes in hourly electricity consumption correlated with the current electricity consumption between $-0.432 < r_i < 0.639$ (see Table II). As with daily rates, the topic that Granger caused the most decrease in power included unigrams such as tweetmyjob and "sonyjob."

E. *Validation Steps*

With Bonferroni correction for multiple tests, we determined the corrected value for $\alpha = 0.05$ to be $\alpha' = \alpha/100 = 0.0005$. Twenty correlations are found to be significant at this rate (see Table II). While we cannot make any explicit claims about the topics this citation [13] determined to have significant relations to power usage, it has been argued [9], [13], [17], [18] that the most common words in a topic are representative of the inherit meaning of the topic. Here, we present the most significant words for each topic, with select words bolded for easier interpretation. With this interpretation in mind, it appears that the three most negatively correlated topics include activity such as having a job, posting on Foursquare or Instagram (i.e., things done outside the residence) and job searches. The top three positively correlated topics include topics about Christmas, storms, and surprisingly, a topic consisting of several vulgarities.

We found a total of 20 statistically significant correlations between events (as inferred by detected topics) and power consumption. Earlier, we presented the 20 topics that had statistically significant correlations with power consumption (see Table II). However, it is also important to consider topics that are rated with a low coefficient of determination to see if

TABLE III
PROBABILITY OF A TOPIC INDEPENDENT AND DEPENDENT ON A POTENTIALLY RELATED EVENT

Topic	$p(Topic)$	$p(Topic\|Event)$	$p(Event\|Topic)$
Sunday	0.00350	0.0182	0.728
Christmas	0.00243	0.256	0.351
Rain	0.0024	0.0137	0.627

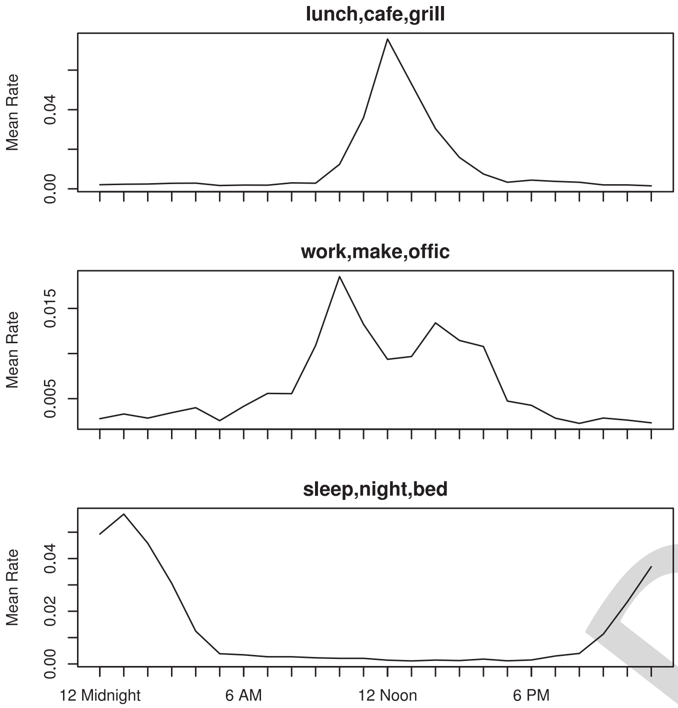

Fig. 5. Mean hourly rate of three select topics. Chart titles indicate three representative words for each topic.

TABLE IV
TOPICS GENERATED THROUGH A REVIEW OF THE LITERATURE, RANKED BY OCCURRENCE IN "NEW & USA" PAPERS

Topic	New & in USA	New	USA
Temperature	4	6	5
Income	3	4	4
Electric Price	3	4	4
Air Conditioner	2	4	5
Heater	2	2	5
Dishwasher	1	2	4
Clothes Dryer	1	2	4
Refrigerator	1	1	2
Water Heater	1	1	3
Building Codes	1	1	1
Own Pool	1	1	1
Own Spa	1	1	1
Lighting	1	1	1
Stove	0	0	3
Freezer	0	0	3
Television	0	0	2
Clothes washer	0	0	1
Wind	0	2	1
Rain	0	1	0
Household Size	0	1	0
Total Papers	7	10	10

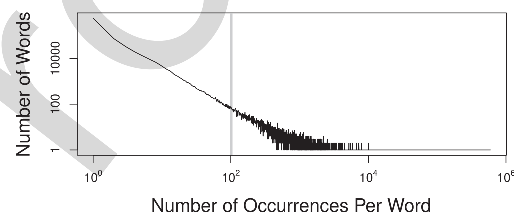

Fig. 6. Distribution of unigrams detected shows a long-tail distribution. The gray line represents the automatically determined cut, w.

they are actually *not* likely to related to residential electricity consumption. The least related topic's three most representative words are "asiathegreat," "manufactur" and "deal." It would appear that these topics are about manufacturing–perhaps in China–which does not have a direct effect on *residential* electricity consumption. The second least related topic's three most representative words are "louisseandon," "ya," and "blo." The third least related contains "justinbieb," "lt," and "sagesummit." These two topics would seem to be related to news about entertainers Louis Sean Don and Justin Bieber, which are likely related to entertainment news rather than electricity consumption.

V. EXPERIMENTS AND RESULTS

One may ask "what is the value of this system over traditional keyword mining or just using expert knowledge?" While our system allows knowledge discovery with limited need for expert knowledge, if it does not perform well, then it is not useful. To justify our system's existence, we compare the results of our system to topics common in the power consumption literature. Additionally, we perform keyword mining to detect words, instead of topics, that are related to electricity consumption.

A. Comparison to Domain Experts

To approximate that knowledge of an expert on power consumption modeling, we perform a literature review. We sample Google Scholar for 100 papers that appear relevant to our question. We discard 85 papers which are either inaccessible (e.g., out of print papers from the '70s), irrelevant to our topic (e.g., a paper on building the Nigerian power grid) or do not explicitly state activities to model (e.g., a paper on synchronizing houses on a smart grid which filter out the customers activities). While we could read the papers for other ideas of important topics, we avoid to because: 1) we risk biasing the set of topics due to selective reading; 2) if a topic is not explicitly modeled or measured, we can assume that the expert does not consider it important; and 3) this literature review is not designed to collect all relevant topics, just ones that are common amongst experts.

Additionally, we separate papers that are more than 10 years old or do not focus on American populations. While these papers may contain expert knowledge, our Twitter and power datasets are based on recent, American usage, which may be different from older usage patterns or those of citizens of other countries. In total, we find 12 topics from recent and local papers [30], [31], [33], [34], [49]–[51] and an additional eight topics from other papers [32], [35], [52]–[57] (see Table IV). Topics were explicitly presented from the papers by either tables or equations. If we only consider the topics that occur more than once in the set of recent and local papers ("temperature," "income," "electricity price," "air conditioner," and "heater"), then we can informally detect two clusters of topics: 1) "climate control" and 2) "economic factors." Both of these two topics were also discovered to be significant measures of electric consumption through our automated system.

Our system found 20 topics that are related to electricity consumption. Our literature review also found 20 topics that are related to electricity consumption. It would seem, however, that these two methods of knowledge discovery discovered topics that were different from each other. The literature review found topics such as temperature or dishwasher usage as interesting topics (see Table IV) while the topic modeling found topics such as having a hangover on the weekend or going to the mall as interesting topics (see Table I). This can be explained by the methods used to collect data. The literature focuses on things that are easy to measure by traditional sensors. However, we use humans as "organic" sensors. This results in different types of data collected: it is easy to have a person report that they are going out on the weekend, but relatively hard to design a sensor to measure this. On the other hand, a sensor to measure temperature is trivial to acquire, but it is unlikely for a person to accurately report the temperature on a regular basis. By focusing on the human element, we have been able to detect important factors of electricity consumption that were previously overlooked due to limitations in traditional sensors and domain knowledge.

Often times, the elements which can easily be studied by these experts and events which are present on social media do not have many commonalities. Discovering these latent events, processed by human sensors, is one major advantage of this paper over traditional sensors. For example, humans might aid in discovering a third variable at work (such as a football game), which leads to an increase in power consumption, while a more guided approach will tend to be informed instead by a television. This demonstrates that not only can we reproduce previous results, but we can also generate novel hypotheses, as told by human sensors.

B. Comparison to Keyword Analysis

We also consider algorithmically generating keywords instead of topics. First the text is cleaned through stemming and *stop word* removal, equivalent to the methods implemented in our system (see Section III-A). Instead of using topic modeling to filter out irrelevant keywords, we are limited to just selecting keywords based on their frequency in the dataset. The $n = 1, 2, \ldots, 5000$ most commonly occurring keywords are selected. The keywords are then tested for relations through cross correlation with the electricity consumption data, the same way that topics were tested for relations in Sections III-D and III-E. We try different values of n because if we try too few keywords, important keywords will be lost, but if we try too many keywords, then, once Bonferroni correction is applied, there will not be enough statistical power to detect significant keywords.

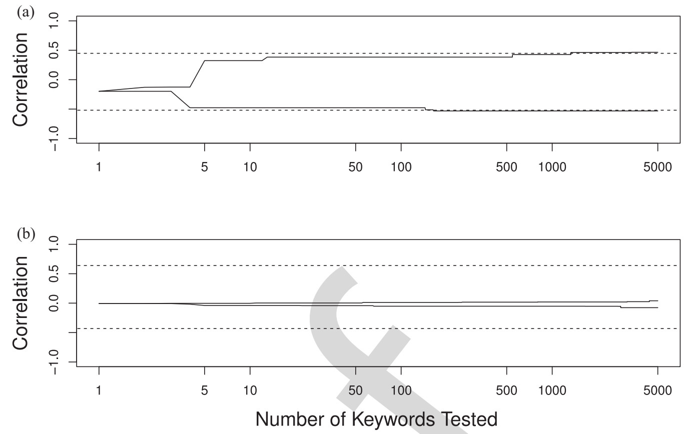

Fig. 7. Strongest positive or negative keyword given a set number of keywords tested. Dashed lines indicate the strongest positive or negative topic detected. Data was aggregated by (a) day or (b) hour.

Additionally, we could define words that occur very frequently in our dataset as de-facto stop words and remove them in addition to the predefined stop word list. However, we do not do this as the tests in this section are independent of each other (besides the Bonferroni correction), compared to the frequency-based methods of our proposed event inference system, so the gain in statistical power is limited in comparison of the risk of removing strongly predictive keywords. Finally, we consider the strongest positive and negative rates of correlation detected for each value of n (see Fig. 7). All minimum and maximum correlations displayed are significant at the 0.05 level, even when Bonferroni correction is applied.

Testing keywords instead of topics resulted in some correlations when dealing with daily aggregation. However, our keyword test allows for a number of tests equivalent to the size of the corpus, which is hard to directly compare against testing 100 topics. When we only consider the top 100 keywords, we find keywords with the strongest positive correlation to be "don" with $r = 0.384$ and the keywords with the strongest negative correlation to be sq with $r = -0.476$. Our system finds events where the strongest positive correlation is 0.448 and the strongest negative correlation of -0.519, a 16.7% and 9.03% improvement, respectively. While keyword-based models do provide some information for daily prediction, hourly prediction does not seem well suited for keyword analysis with correlations ranging between -0.074 and 0.004, limiting the usefulness of previous methods for fine-grained prediction. Comparatively, our system which finds topics that match power usage with correlations between -0.432 and 0.639 resulting in an increase of explained variance of up to 41%.

VI. Predicting Future Electrical Consumption

Up to this point we have only considered individual topics to predict the phenomena. Here, we consider multivariable regression based on lagged predictive variables to predict hourly power usage (see Algorithm 5). As a baseline, we consider a 12-variable auto-correlation model where the maximum lag of 12 was determined through maximum likelihood estimation. We then compare this model to

TABLE V
CORRELATION COEFFICIENTS FOR MODELS USING AUTO-CORRELATION, TOPICS, OR A SUBSET OF ATTRIBUTES

	Auto-Corr	Topics	Auto-Coor + Topics	Subset
Training Set	0.9515	0.9430	0.9788	0.9777
5-fold CV	0.9510	0.9116	0.9670	0.9682
80%/20%	0.9313	0.7152	0.9003	0.9632

TABLE VI
ROOT MEAN SQUARE ERRORS FOR MODELS USING AUTO-CORRELATION, TOPICS, OR A SUBSET OF ATTRIBUTES

	Auto-Corr	Topics	Auto-Coor + Topics	Subset
Training Set	39.6508	42.9102	26.3846	27.0747
5-fold CV	39.8758	53.2473	32.8872	32.2713
80%/20%	51.7108	121.166	66.3104	34.9691

three models: a multivariable regression on the detected topics, a multivariable regression on the 38 topics that were found to have a Granger causal relationship to electricity consumption *and* the auto-correlation model, and the second model with a subset of the attributes used. Which attributes are retained in the third model are selected through removing attributes with the smallest coefficients and refitting the model until AIC no longer improves.

AQ3

We now determine the accuracy of each model by determining the correlation coefficient for either through traditional statistical methods, fivefold cross validation, or a 80%/20% test-train split. The 80%/20% test-train split is performed on data that is ordered by time where the fivefold cross validation is performed on randomly ordered data. We find that at least one of our models out perform the base-line in all three evaluation methods. Importantly, the 80%/20% test-train split represents the most realistic case of predicting future electricity usage, and our model provides an additional 4.28% explanation of electricity usage. These results can be seen in Tables V and VI.

A. Comparison With U.S. DOE Model

The U.S. Department of Energy provides Commercial and Residential hourly load profiles for typical meteorological year (TMY3) locations around the United States. These simulated values are derived from a combination of weather data from the National Solar Radiation Database,[5] regional climate-specific information (cold/very cold, hot-dry/mixed-dry, hot-humid, marine, and mixed-humid), and load profile type (high, base, and low) which define physical building characteristics such as home size, layout, insulation type, heating fuel source, and occupants. These simulations take into account very detailed electricity demands, (e.g., heat output by showers and dishwasher temperature point) and provide an hourly demand of an average household in each of hundreds of sites around the United States. Incorporating all of this information, this model presents a year-agnostic estimation of the hourly electricity usage of households across the country. That is, the model does not differentiate between 1 A.M., January 1, 2011, and 1 A.M. January 1, 2012. Rather, it assumes each hour is the same. The DOE has made this model publicly available for researchers seeking to predict energy demands across U.S. Cities.[6]

[5]https://mapsbeta.nrel.gov/nsrdb-viewer

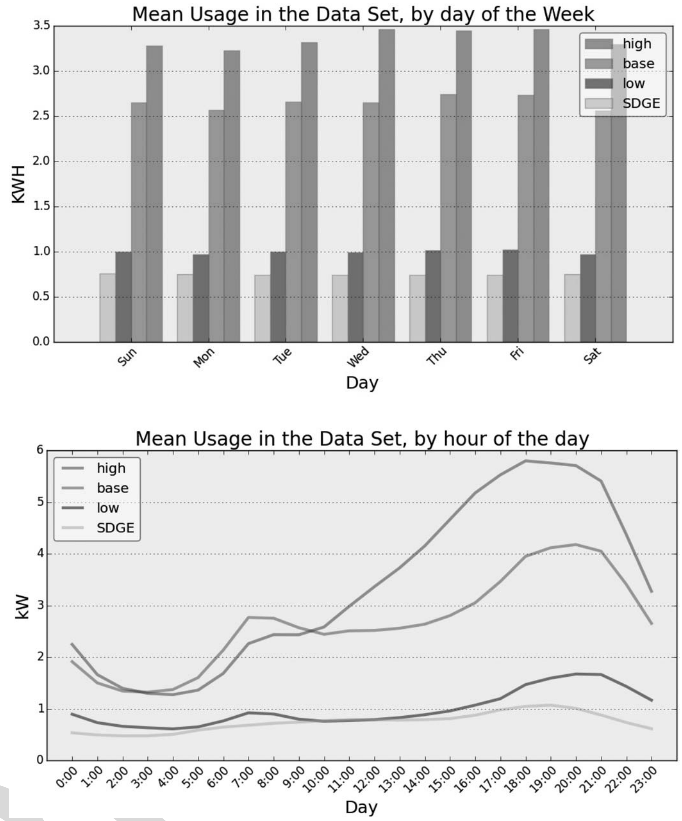

Fig. 8. Periodicity of SDGE provided energy data, compared to TMY3 simulated data.

To test the efficacy of the TMY3 models in simulating the real world energy use of the San Diego area, we compared the TMY hourly use with the SDGE-provided data from Section IV. The TMY3 data is considered the baseline model, with the SDGE data representing the ground truth. Since the TMY3 data is year agnostic, variations in energy use due to severe weather events (as opposed to seasonality), and date-specific periodicity (weekends and weekdays) will not be included. These differences can be seen in Fig. 8. While the SDGE data is lower in magnitude than the TMY3 load profiles, the general trends of the data are reflected best by the *base* model, which carries an hourly correlation coefficient of 0.7544 and an RMSE of 130 when used as input for a linear regression of the SDGE data.

Next, TMY3 data is used to predict monthly SDGE electricity usage. The monthly usage data is provided by SDGE, aggregated across customers in each zip code.[7] This data is shown in Fig. 9. Note that since the TMY3 is year agnostic, the data will repeat on an annual cycle. Once again, the magnitude of each of the load models is higher than the aggregate data provided. When analyzed against the real monthly data for San Diego homes, no single model consistently correlates better than the others, with the *high* model performing best

[6]http://en.openei.org/datasets/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states
[7]https://energydata.sdge.com/

TABLE VII
ρ AND RMSE FOR EACH TMY MODEL

year	ρ			RMSE		
	high	base	low	high	base	low
2012	0.65	0.21	0.59	121.4	156.3	129.3
2013	0.58	0.81	0.79	63.6	45.5	47.5
2014	0.82	0.78	0.93	40.7	45.1	27.2
Aggregated	0.61	0.43	0.64	83.5	94.8	80.1

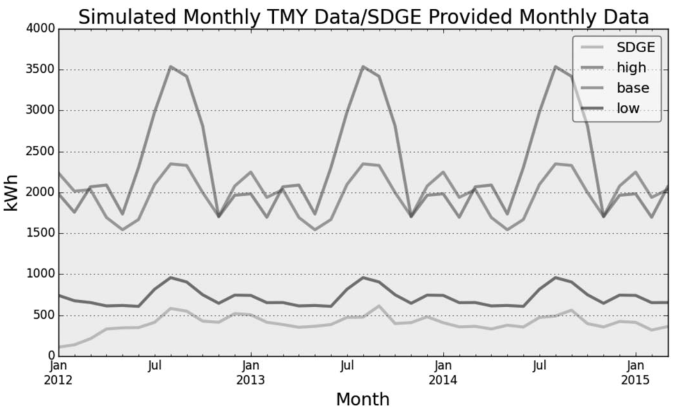

Fig. 9. TMY3 data, aggregated by month, compared with SDGE monthly data.

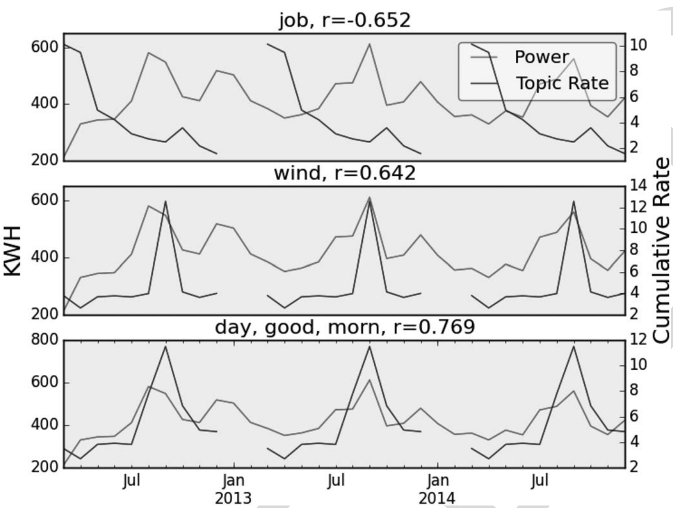

Fig. 10. Topic rates for three sample topics. Note the recurrence of the topic rate, as the topics were analyzed for 1 year only.

in 2012, *base* in 2013, and *low* in 2014. These same models possess the lowest RMSE on a yearly basis, as seen in Table VII.

Finally, we demonstrate that our proposed social media model outperforms the TMY3 model, given the same ground truth (SDGE data), by using the topic models and frequencies from Sections IV–VI. As with the TMY3 data, we assumed that each topic frequency is repeated for that same hour and date on all subsequent years. Similar to Fig. 4, these cumulative topic rates by month can be seen in Fig. 10. Next, these topics were aggregated on a monthly basis, the significance of each topic was tested, and the Bonferroni correction applied, leaving 13 topics whose $p < 0.05/100$. Finally, we used these frequencies as input in a regression model for March–December of each year. This model yielded an RMSE of 43.6 when applied to this time period, which outperforms the linear regression performance of the best TMY3 data in Table VII, whose best models RMSE was 80.1, an 83%.

VII. CONCLUSION

In this paper, we proposed a theoretical backing to our design (see Section III), which assumed a link between: 1) events and text; 2) text and word vectors; 3) word vectors and topics; 4) topics and events; and 5) events and real-world phenomena. We now provide evidence of these relations. Previous work [9], [39] has verified that events cause users to post on social media networks. Similarly, the conversion of text into word vectors has previously been discussed [4], [17], [20], [41], [42]. The most likely words are cohesive within each topic and have large between-topic variation (see Table I). Thus it is likely that topics can be generated from social media network text using LDA [14], [15]. We choose three topics that contain words related to Sundays, Christmas, and storms. By studying the temporal patterns of each topic, we find a relationship between the storm topic and the days with “rain” events in San Diego, the Sunday topic to be most often discussed on Sundays, and the Christmas topic to trend during December (see Fig. 4). Finally, we show a relationship between our discovered events and energy consumption through statistical analysis (see Table II). Hence, we conclude that there is evidence for our assumptions on links, at least when applied to our case study.

We presented a novel form of semi-supervised knowledge discovery that infers events from topics generated from social media network data. These events are then used to form hypotheses about real-world phenomena which are then validated. To provide support for our case, we perform a case study where Twitter data is used to predict electricity consumption rates. The results are then compared to topics generated by domain experts and keyword analysis. We find that our system detects events tangential to what the literature is currently focused on and that our system outperforms an equivalent keyword analysis by up to 16.7%. When combined with time-series modeling, we are able to predict electricity consumption with correlations of up to 0.9788 and a mean absolute error of 19.84 watts—less than the energy consumption of a single light bulb. Finally, we compared the performance of this model to the models generated by the DOE for the San Diego area, and found it to be more accurate.

Future work may consider a more robust comparison of this model against other existing models, since several such models exist. Additionally, this model might be employed for a more directed event detection, as described in the introduction. The textual analysis in this paper could be augmented by considering synonyms and related concepts through word embedding which groups similar words together automatically. Additionally, other data modalities might also be considered, such as images, videos, and social media metadata. Since there is a spatial component of this data, future work may also analyze similar data for a different part of the country, to determine if the trends we have identified hold true elsewhere. Finally, it may prove fruitful to analyze a similar methodology for other utilities such as water.

ACKNOWLEDGMENT

Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors. The views in this document are those of the authors and do not reflect the official policy or position of the U.S. Air Force, Department of Defense, or the U.S. Government.

REFERENCES

[1] L. Zhao, S. Sakr, A. Liu, and A. Bouguettaya, *Cloud Data Management*. New York, NY, USA: Springer 2014. AQ4

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding Twitter data with TweetXplorer," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Chicago, IL, USA, 2013, pp. 1482–1485. [Online]. Available: http://doi.acm.org/10.1145/2487575.2487703

[4] T. Bodnar, V. C. Barclay, N. Ram, C. S. Tucker, and M. Salathé, "On the ground validation of Online diagnosis with Twitter and medical records," in *Proc. 23rd Int. Conf. World Wide Web Companion*, Seoul, South Korea, 2014, pp. 651–656.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[6] M. Eirinaki, M. D. Louta, and I. Varlamis, "A trust-aware system for personalized user recommendations in social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 4, pp. 409–421, Apr. 2014.

[7] M. J. Lanham, G. P. Morgan, and K. M. Carley, "Social network modeling and agent-based simulation in support of crisis de-escalation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 1, pp. 103–110, Jan. 2014.

[8] T. Bodnar, C. Tucker, K. Hopkinson, and S. G. Bilén, "Increasing the veracity of event detection on social media networks through user trust modeling," in *Proc. IEEE Big Data*, Washington, DC, USA, 2014, pp. 636–643.

[9] D. D. Ghosh and R. Guha, "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system," *Cartography Geogr. Inf. Sci.*, vol. 40, no. 2, pp. 90–102, 2013.

[10] A. Smith and J. Brenner, *Twitter Use 2012*. Pew Internet & Amer. Life Project, 2012. AQ5

[11] T. Bodnar and M. Salathé, "Validating models for disease detection using Twitter," in *Proc. 22nd Int. Conf. World Wide Web Companion*, Rio de Janeiro, Brazil, 2013, pp. 699–702.

[12] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, "Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales," *PLoS Comput. Biol.*, vol. 9, no. 10, 2013, Art. no. e1003256.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[14] H. D. Kim *et al.*, "Mining causal topics in text data: Iterative topic modeling with time series feedback," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 885–890.

[15] X. W. Zhao, J. Wang, Y. He, J.-Y. Nie, and X. Li, "Originator or propagator?: Incorporating social role theory into topic models for twitter content analysis," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 1649–1654.

[16] M. Wahabzada, K. Kersting, A. Pilz, and C. Bauckhage, "More influence means less work: Fast latent Dirichlet allocation by influence scheduling," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, Glasgow, U.K., 2011, pp. 2273–2276.

[17] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & Web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, Beijing, China, 2008, pp. 91–100.

[18] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5228–5235, Apr. 2004.

[19] I. Bíró, J. Szabó, and A. A. Benczúr, "Latent Dirichlet allocation in Web spam filtering," in *Proc. 4th Int. Workshop Adversarial Inf. Retrieval Web*, Beijing, China, 2008, pp. 29–32.

[20] J.-C. Guo, B.-L. Lu, Z. Li, and L. Zhang, "Logisticlda: Regularizing latent Dirichlet allocation by logistic regression," in *Proc. PACLIC*, Hong Kong, 2009, pp. 160–169.

[21] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram, "Discovering health-related knowledge in social media using ensembles of heterogeneous features," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 1685–1690.

[22] L. Dannecker *et al.*, "pEDM: Online-forecasting for smart energy analytics," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manag.*, San Francisco, CA, USA, 2013, pp. 2411–2416.

[23] V. Aravinthan, V. Namboodiri, S. Sunku, and W. Jewell, "Wireless AMI application and security for controlled home area networks," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Detroit, MI, USA, Jul. 2011, pp. 1–8.

[24] C. Bennett and D. Highfill, "Networking AMI smart meters," in *Proc. IEEE Energy 2030 Conf. ENERGY*, Atlanta, GA, USA, Nov. 2008, pp. 1–8.

[25] C. Bennett and S. B. Wicker, "Decreased time delay and security enhancement recommendations for AMI smart meter networks," in *Proc. Innov. Smart Grid Technol. (ISGT)*, Gaithersburg, MD, USA, Jan. 2010, pp. 1–6.

[26] J. E. Fadul, K. M. Hopkinson, T. R. Andel, and C. A. Sheffield, "A trust-management toolkit for smart-grid protection systems," *IEEE Trans. Power Del.*, vol. 29, no. 4, pp. 1768–1779, Aug. 2014. AQ6

[27] M. T. O. Amanullah, A. Kalam, and A. Zayegh, "Network security vulnerabilities in SCADA and EMS," in *Proc. IEEE/PES Transm. Distrib. Conf. Exhibit. Asia Pac.*, Dalian, China, 2005, pp. 1–6.

[28] P. Palensky, E. Widl, and A. Elsheikh, "Simulating cyber-physical energy systems: Challenges, tools and methods," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 3, pp. 318–326, Mar. 2014.

[29] A. Aroonruengsawat, M. Auffhammer, and A. H. Sanstad, "The impact of state level building codes on residential electricity consumption," *Energy J.*, vol. 33, no. 1, pp. 31–52, 2012.

[30] I. Ayres, S. Raseman, and A. Shih, "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage," *J. Law Econ. Org.*, vol. 29, no. 5, pp. 992–1022, 2013.

[31] I. M. L. Azevedo, M. G. Morgan, and L. Lave, "Residential and regional electricity consumption in the U.S. and EU: How much will higher prices reduce CO_2 emissions?" *Elect. J.*, vol. 24, no. 1, pp. 21–29, 2011.

[32] M. Filippini, "Short- and long-run time-of-use price elasticities in Swiss residential electricity demand," *Energy Policy*, vol. 39, no. 10, pp. 5811–5817, 2011.

[33] D. Livengood and R. Larson, "The energy box: Locally automated optimal control of residential electricity usage," *Service Sci.*, vol. 1, no. 1, pp. 1–16, 2009.

[34] D. Petersen, J. Steele, and J. Wilkerson, "WattBot: A residential electricity monitoring and feedback system," in *Proc. Extended Abstracts Human Factors Comput. Syst. (CHI)*, Boston, MA, USA, 2009, pp. 2847–2852.

[35] K. Wangpattarapong, S. Maneewan, N. Ketjoy, and W. Rakwichian, "The impacts of climatic and economic factors on residential electricity consumption of Bangkok Metropolis," *Energy Build.*, vol. 40, no. 8, pp. 1419–1425, 2008.

[36] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Proc. Workshop Data Min. Appl. Sustain. (SIGKDD)*, San Diego, CA, USA, 2011. AQ7

[37] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Secur. Privacy*, vol. 8, no. 1, pp. 11–20, Jan./Feb. 2010.

[38] S. Wicker and R. Thomas, "A privacy-aware architecture for demand response systems," in *Proc. 44th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Kauai, HI, USA, 2011, pp. 1–9.

[39] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol, CA, USA: O'Reilly, 2013.

[40] M. F. Porter, "An algorithm for suffix stripping," *Program Electron. Library Inf. Syst.*, vol. 14, no. 3, pp. 130–137, 1980.

[41] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Berkeley, CA, USA, 1999, pp. 42–49.

[42] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, 2004. AQ8

[43] M.-P. Kwan, "The uncertain geographic context problem," *Ann. Assoc. Amer. Geographers*, vol. 102, no. 5, pp. 958–968, 2012.

[44] D. W. S. Wong, "The modifiable areal unit problem (MAUP)," in *WorldMinds: Geographical Perspectives on 100 Problems*. Dordrecht, The Netherlands: Springer, 2004, pp. 571–575.

[45] G. Heinrich, "Parameter estimation for text analysis," Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2005. AQ9

[46] C. W. J. Granger, "Some recent development in a concept of causality," *J. Econometrics*, vol. 39, nos. 1–2, pp. 199–211, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0304407688900450

[47] H. H. Lean and R. Smyth, "Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia," *Energy*, vol. 35, no. 9, pp. 3640–3648, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544210002719

[48] R. J. Cabin and R. J. Mitchell, "To Bonferroni or not to Bonferroni: When and how are the questions," *Bull. Ecol. Soc. Ameri.*, vol. 81, no. 3, pp. 246–248, 2000.

[49] A. Aroonruengsawat and M. Auffhammer, *Impacts of Climate Change on Residential Electricity Consumption: Evidence From Billing Data*. Chicago, IL, USA: Univ. Chicago Press, 2011.

[50] A. Jacobson, A. D. Milman, and D. M. Kammen, "Letting the (energy) Gini out of the bottle: Lorenz curves of cumulative electricity consumption and Gini coefficients as metrics of energy distribution and equity," *Energy Pol.*, vol. 33, no. 14, pp. 1825–1832, 2005.

[51] S. Kishore and L. V. Snyder, "Control mechanisms for residential electricity demand in smartgrids," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Gaithersburg, MD, USA, 2010, pp. 443–448.

[52] K. P. Anderson, "Residential energy use: An econometric analysis," RAND, Santa Monica, CA, USA, Tech. Rep. R-1297-NSF, Oct. 1973. AQ10

[53] D. W. Caves and L. R. Christensen, "Econometric analysis of residential time-of-use electricity pricing experiments," *J. Econ.*, vol. 14, no. 3, pp. 287–306, 1980.

[54] J. K. Dobson and J. D. A. Griffin, "Conservation effect of immediate electricity cost feedback on residential consumption behavior," in *Proc. 7th ACEEE Summer Study Energy Efficiency Build.*, vol. 2. Pacific Grove, CA, USA, 1992, pp. 33–35.

[55] G. Lafrance and D. Perron, "Evolution of residential electricity demand by end-use in quebec 1979-1989: A conditional demand analysis," *Energy Stud. Rev.*, vol. 6, no. 2, pp. 164–173, 1994.

[56] I. Matsukawa, "The effects of information on residential demand for electricity," *Energy J.*, vol. 25, no. 1, pp. 1–17, 2004.

[57] M. Parti and C. Parti, "The total and appliance-specific conditional demand for electricity in the household sector," *Bell J. Econ.*, vol. 11, no. 1, pp. 309–321, 1980.

AQ11 AQ12

Todd Bodnar (M'XX) received the B.Sc. degree in computer science from the Pennsylvania State University, State College, PA, USA, in 2012, and the Ph.D. degree in biology in 2015.

His current research interests include machine learning and data mining on large datasets to measure sociological patterns.

Matthew L. Dering received the B.A. degree in psychology from Swarthmore College, Swarthmore, PA, USA, in 2007, and the M.S. degree in computer science from the Pennsylvania State University, State College, PA, USA, in 2014, where he is currently pursuing the Doctoral degree under the supervision of Dr. C. Tucker.

His research interests include computer vision, novel data sources, and video analysis, especially pertaining to sports.

Conrad Tucker (M'XX) received the B.S. degree in mechanical engineering from the Rose-Hulman Institute of Technology, Terre Haute, IN, USA, in 2004, and the M.S. degree in industrial engineering, the M.B.A. degree in business administration, and the Ph.D. degree in industrial engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA.

His current research interests include formalizing system design processes under the paradigm of knowledge discovery, optimization, data mining, informatics, applications in social media network mining of complex systems, design, and operation, product portfolio/family design, and sustainable system design optimization in the areas of energy, healthcare, consumer electronics, environment, and national security.

Kenneth M. Hopkinson (SM'XX) received the B.S. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1997, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, USA, in 2002 and 2004, respectively, all in computer science.

He is a Professor of Computer Science with the Air Force Institute of Technology, Wright-Patterson AFB, OH, USA. His current research interests include simulation, networking, and distributed systems.

AUTHOR QUERIES
AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

If you have not completed your electronic copyright form (ECF) and payment option please return to the Scholar One "Transfer Center." In the Transfer Center you will click on "Manuscripts with Decisions" link. You will see your article details and under the "Actions" column click "Transfer Copyright." From the ECF it will direct you to the payment portal to select your payment options and then return to ECF for copyright submission.

AQ1: Fig. 6 is cited before Fig. 4. Please check if Fig. 6 can be renumbered so that they are cited in sequential order.
AQ2: Please cite "Fig. 2" and "Table III" inside the text.
AQ3: Please provide expansion for the term "AIC."
AQ4: Please confirm that the location and publisher information for References [1] and [44] is correct as set.
AQ5: Please provide the location for Reference [10].
AQ6: Please confirm that the edits made to Reference [26] is correct as set.
AQ7: Please provide the page range for Reference [36].
AQ8: Please provide the issue number or month for Reference [42].
AQ9: Please provide the department name and technical report number for Reference [45].
AQ10: Please provide the department name for Reference [52].
AQ11: Please provide the membership years for the authors "T. Bodnar, C. Tucker, and K. M. Hopkinson."
AQ12: Please provide the organization name for the degrees attained by the author "T. Bodnar."