2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Twitter Social Contagion Monitor

Vladimir Barash
Graphika Labs
Graphika, Inc.
New York, USA
vlad.barash@graphika.com

Clayton Fink
Applied Physics Laboratory
Johns Hopkins University
Maryland, USA
Clayton.Fink@jhuapl.edu

Christopher Cameron
Social Dynamics Lab
Cornell University
Ithaca, USA
cjc73@cornell.edu

Aurora Schmidt
Applied Physics Laboratory
Johns Hopkins University
Maryland, USA
Aurora.Schmidt@jhuapl.edu

Wei Dong
Social Dynamics Lab
Cornell University
Ithaca, USA
wd97@cornell.edu

Michael Macy
Social Dynamics Lab
Cornell University
Ithaca, USA
mwmacy@cornell.edu

John Kelly
Graphika Labs
Graphika, Inc.
New York, USA
john.kelly@graphika.com

Amruta Deshpande
Graphika Labs
Graphika, Inc.
New York, USA
amruta.deshpande@graphika.com

***Abstract*—We describe and validate a system for monitoring social contagions on Twitter: social movements, rumors, and emotional outbursts that spread from person to person in a viral manner. We use Twitter streams to monitor the spread of these phenomena through human social and information networks. This system, the contagion monitor, parses Twitter posts to identify emerging phenomena, as captured in hashtags, URLs, words and phrases, or account-handles, and then determines the extent to which a particular phenomenon spreads via the social network (in contrast to its spread via news broadcasts or independent adoption) and locates the contagion within Twitter communities. The monitor approximates the adoption threshold of a social contagion by measuring the fraction of Twitter users who were "infected" by the contagion (e.g., joined a particular social movement) after more than one of their friends had done so. Finally, the monitor makes a judgment about whether the phenomenon has reached critical mass, which is defined as the point where a social contagion begins spreading rapidly and breaches the social boundaries of its early adopter group. We test our prototype monitor on two data sources — an ongoing stream of tweets grouped by user-added hashtags and a collection of posts by a monitored set of Nigerian Twitter users — before productionalizing. We use the Amazon Mechanical Turk platform to evaluate the performance on both data sources. In both cases, we find that our approach successfully distinguishes between high-threshold and low-threshold social contagions.**

I. Introduction

The past decade has seen transformative social movements, such as the Arab Spring, Black Lives Matter, and the movements which led to elections of Barack Obama and Donald Trump to the US presidency. Recent approaches in social science [1]–[6] seek to understand the dynamics by which social movements break out of local contexts to become widespread phenomena. We leverage these approaches to construct an automated tool, the contagion monitor, to detect emerging social movements before they gain widespread popularity. Our tool expressly seeks out the more fundamental and transformative social movements in social media, capable of bringing about real behavior change, and in so doing, offers an advantage over other approaches (see section II-C) which look for trends. We tested this tool in two large-scale empirical settings and subsequently deployed it in a productionalized application that is currently in daily use.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

Our innovation relies on a network structural approach that can distinguish between transient movements (e.g., viral memes), of low social impact, and more transformative ones (e.g., voting or protesting). We are able to make this distinction based on how information cascades move through social networks, without relying on time-consuming language analysis or contextual searches. We have augmented our tool with network-based metrics, which enable the analyst to monitor social contagions before they become widely relevant.

We define "social contagions" as social movements, rumors, and emotional outbursts that spread from person to person in a viral manner. Our social contagion monitor processes streams of Twitter data to scan for these. For ethical reasons, we avoid running any sort of controlled experiment on large human populations; instead, we treat data collected from the monitor as purely observational or "natural experiment" outcomes.

In scanning for social contagions, the monitor makes two key algorithmic judgments about their dynamics. For each, the tool calculates metrics that estimate the threshold of participation in the movement [7], indicating transformative potential; and whether the movement is about to break out of a local network cluster [1], indicating virality. Together, these judgments can help researchers differentiate between three categories of movements, from the transient to beyond:

a) Low-threshold movements that have low cost to participation or network externality: these can spread quickly, but some scholars (e.g., [7]) speculate they are unlikely to bring about transformative political or social change.
b) High-threshold movements that have not yet broken out of local network clusters: these movements may be important to monitor but do not yet have the reach to bring about transformative political or social change.
c) High-threshold movements that have broken out of local networks: these are both costly (i.e. sufficiently "high

stakes") and have the reach to have a relatively high probability of affecting political or social change.

We tested two setups of the monitor: the regional contagion monitor (RCM) and the streaming contagion monitor (SCM). These implementations differ by the type of data they consume and the criteria for detecting emerging phenomena. The RCM focuses on a smaller geographic region and employs significant historical data for detecting new contagions while the SCM processes streaming data with minimal use of historical data.

With the RCM, we test our system in a more controlled setting of a fixed, more narrowly scoped community of regional Twitter users over long time. Here, we can determine which diffusion events - captured via hashtags, URLs, or words and phrases - are propagating according to theoretical expectations. If productionalized, this setup would offer regionally focused insights into emerging movements and their transformative potentials. With the SCM, we extend the scope of the monitor to a wider range of sociocultural settings, dynamically capturing the engaged network and allowing the setup to pick out more recent, potentially transformative world-wide movements. The trade-off between these setups is one of greater depth (RCM) versus breadth (SCM) of topics and regions. We have currently productionalized the SCM setup for daily use.

Next, we describe prior theoretical work and other social media trend tools. In section III we describe the monitoring systems including what data they ingest and some risks. In section IV, we describe a framework for evaluating the monitors' performance embedded within two science questions and share those results. And finally, we conclude in section V with a discussion of results and some future work.

II. Background

A. Social Movements and Contagion Models

Reference [8] formulated a model of threshold-based adoption cascades on populations in lattice networks, where the threshold depends on k, the number of network neighbors of a node. Reference [7] extended [8]'s model to the Small World [9] network model, which involves the random rewiring of ties on lattice networks; a closer approximation to real social networks. Reference [9] tested the spread of simulated simple and complex contagions on Small World networks. Simple contagions, defined by their threshold of $1/k$, model the spread of disease, information, and easily adopted behaviors that do not require a confirmatory (and redundant) exposure to become infected. In contrast, complex contagions, defined to have have thresholds greater than $1/k$, model behaviors like protesting, rioting, or adoption of new conventions or technologies that carry a greater cost to adoption and require more than one confirmatory exposure prior to infection.

Reference [7] found that in Small World networks, complex contagions rarely spread beyond the initial "seed" cluster, except occasionally when these can leverage shortcuts that result from re-wiring to new clusters. Typically, the re-wiring reduces the confirmatory exposures available to the complex contagion to spread. Reference [1] discovered a *critical mass* point in the fraction of infected nodes beyond which the contagion spreads through the full network quickly and with high probability (like simple contagions). Beyond the critical mass, the chance of sufficient exposures to leverage shortcuts to new clusters becomes very high.

The authors in [1] confirm that complex contagions require a dense seed cluster in the initial stages of propagation to reach critical mass. They found that the critical mass point has a two-part statistical signature: 1) the contagion's rate of propagation dramatically changes from negative (as the contagion begins to saturate the local seed cluster) to positive (as the contagion begins to spread in fresh network regions) and 2) there is a sharp drop in the density of the network neighbors of new adopters. The first indicates that the contagion has broken out of the seed cluster while the second indicates that it is growing in an unsaturated region. Together, these mean that the contagion is leveraging the randomly re-wired long-range ties and is not limited by the local structure where it begins.

B. Social Movements and Social Media

Many researchers have attempted to study viral behavior in general, and complex and simple contagions in particular, in online social network data. Work by [5] observed that politically-themed hashtags behave like complex contagions, whereas hashtags corresponding to neologisms and Twitter idioms behave like simple contagions. Reference [4] found evidence of social influence and complex contagions in Twitter recruitment networks around social mobilization in Spain in May 2011. Reference [10] compiled a large overview of viral messages, including those related to social contagions, and pointed to the important role for "gatekeepers" whose influence can cause a social contagion to go viral. Reference [2] found the hashtag #bringbackourgirls, relating to the movement to bring back hundreds of girls kidnapped by Boko Haram in Nigeria in 2014, resembled a complex contagion.

Reference [6] found that the diffusion of campaign donations is a complex contagion driven by independent social reinforcement. The authors found that people are more likely to donate if exposed to donors from different social groups than equally many donors from the same group, which suggests an important extension of the complex contagion model: high threshold contagions may require not just multiple sources of exposure but multiple independent sources. Accordingly, we equipped our contagion monitor to be able to identify the distribution of adopters across social groups.

C. Social Media Monitoring Tools

There are a number of social media monitoring tools, both in academic [11], [12] and industrial [13]–[15] settings, and our monitor offers some advantages over these. The Observatory on Social Media and NIFTY analyze the spread of information rather than of social contagions specifically. Crimson Hexagon and other industrial tools focus on general monitoring (influencer identification, trends, geo-location) rather than the specific identification of social contagions and analysis of their dynamics as we do. Our metrics, by design, identify high-threshold social contagions, more likely to be of actionable interest. By analyzing the dynamics of information cascades, and using simple related metrics, we offer a faster and independent signal of emerging movements. Our monitor is thus compatible with and goes beyond the capabilities of existing tools.

III. Methods and Data

A. IRB Compliance

Our research received an IRB exemption because we analyze publicly available data that we also de-identify. We have made it a priority to minimize the risk of exposing sensitive personal data, whether through a leak or hack of our data store, or through unwitting exposure via publication.

B. Risks of Collecting Data on Social Movements

Our data collection is associated with two major sources of risk. First, the data we collect are public Twitter posts, which may contain personally identifying information. A Pew internet survey of teenage use of social media [16] shows that 24% of teens used Twitter in 2012. The same survey shows that 91% of teens post a photo of themselves on Facebook, 71% post their school name, and so on, in the same time period. The scale of our data collection is highly likely to result in including personally identifying information on minors, which requires special care to store or process [17].

Our streaming and regional contagion monitors analyze social contagions, which include social movements. The second risk associated with our data collection is that some of these movements (like the pro-LGBT movement in Nigeria) carry severe risks for participation, including imprisonment. The Same Sex Marriage Prohibition Act in Nigeria "imposes a 10-year prison sentence on anyone who "registers, operates or participates in gay clubs, societies and organization" or "supports" the activities of such organizations" [18]. Simply exposing the identities of Nigerian individuals involved in a pro-LGBT hashtag on Twitter could subject such individuals to the effect of this law. We recognize the severe risks of improper data management when constructing the social contagion monitor.

We do not publish tweets from our contagion monitors except as one-off examples, in which case, we blur all identifying information in the tweet text or metadata. In all publications (including this one), we report summary statistics about hashtags and other features identified by the social contagion monitor rather than individual-level behavior.

C. Data Ingestion

1) Streaming Contagion Monitor: The SCM collects streaming, public Twitter posts (or tweets) and extracts network ties from the same to analyze a contagion. In this study, in order to additionally evaluate complex contagion theory from our results, we chose to scope our streaming collection on themes identified in thirty of [14]'s library of maps. These themes span mulitiple sociocultural contexts — US and European politics, industry verticals such as automotive and food, and large sports events. Reference [14] constructs themed maps, or networks of Twitter accounts that engage with themed conversations, and then collects the accounts' live streaming tweet activity[1], which we accessed. Importantly, we note that these data are not filtered geographically or by keyword, and truly represent the conversation at large on Twitter. We deliberately chose both political and non-political maps to capture both high-threshold social movements and low-threshold news events.

The SCM passes these streaming tweets to our sliding window monitor (SWM), which is capable of real-time tweet stream processing. It does 3 main tasks: feature extraction, tracking features, and nominating features of interest. The SWM acts as a front line filter, reducing the volume of data that is passed on to later analysis stages. In this study, the features are based on hashtags alone. For each incoming tweet, the SWM extracts the hashtag data and creates features that represent hashtags, retweeted hashtags, hashtag pairs and retweeted hashtag pairs. Hashtag pairs are sets of hashtags that appear together in the same tweet. The SWM tracks the frequency of each feature over sliding intervals of 10 minutes, 1 hour and 4 hours, reporting any features that exceed a threshold number of appearances. The SWM aggregates the reports to identify the most popular features over the last 24 hours and nominates the top features for further analysis.

2) Regional Contagion Monitor : We invoke the regional contagion monitor, built and studied in [19], as a static network on which to compute contagion measures for comparison against the same measures computed for the SCM. The RCM focused its data collection on networked accounts that were likely to be located in a specific geographic region. In this way, we approximate a physical social network that is static, and measure its social media activity to look for contagions.

Full details of the RCM data collection are available in [19], but we summarize some key elements here. We identified a set of 108,744 Twitter users located in Nigeria[2] between January and mid-April of 2017, that also remained public and active through our later data collection, and for whom we could carefully assess location. We constructed a network of these users by leveraging the "follows" relationship on the platform, and collected approximately 270 million tweets[3] between April and November of 2017. Our final graph, on which we computed contagion measures, consisted of 103,659 nodes, 11,753,606 edges and an average degree of 113.

3) Productionalized Streaming Contagion Monitor: The productionalized contagion monitor setup, which is in use today, is based on the SCM, but is adapted to [14]'s preparsed data stores and collection. It thus bypasses the SWM's input handling, but computes the same metrics for candidate selection from high resolution feature (e.g., hashtags, urls) occurrence time-streams computed from the [14] data.

[1] `statuses/user_timeline` endpoint – https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

[2] https://www.cia.gov/library/publications/the-world-factbook/geos/ni.htm

[3] `statuses/user_timeline` public Twitter API endpoint

D. Candidate Hashtag Selection

1) Streaming Contagion Monitor: The candidate selector collects the most popular features over the last 24 hours, and then applies a multi-step filter and an exclusion-list filter. The multi-step seeks to filter out features that have appeared in the top 300 most popular, at any time in the past 5 days, with one exception. If in the last 5 days, the number of communities identified by [14] in which the feature is relevant [20] has increased by 20%, then it is not filtered out. The exclusion-list filter contains hashtags that are spam or regularly reoccurring Twitter "memes" (e.g. #followfriday) that are obviously not related to social movements.

2) Regional Contagion Monitor: For each calendar day of tweets (assuming midnight as GMT+1 or West African Time), the regional contagion monitor extracted all hashtags used during that day and the 30 days prior to that. For those hashtags used by 100 or more unique users, we classified a hashtag as a candidate for analysis if its count for that day was two standard deviations greater than its mean count for the previous 30 days. Our contagion analysis for a given day was restricted to these hashtags. We excluded from analysis hashtags that were on our exclusion-list (common ones such as #NP) or hashtags used by monitored users in the six months prior to August 1, 2017. The monitor nominated 2,823 trending hashtags from August through October 2017.

E. Contagion Analysis

1) Streaming Contagion Monitor: For each feature identified by the candidate selector, the contagion analyzer collects up to the last five days of tweets with the hashtag. From this, it identifies adopters and constructs an adopter network [2] by connecting users who have retweeted or mentioned (or been retweeted by or mentioned by) any adopter. Then it computes the following metrics for each feature:

a) Cumulative distribution, over k, (abbreviated CDF$_k$) of percent of users who started tweeting about the feature after k or fewer network neighbors had done so.
b) Percent of all users who tweeted the feature before any of their network neighbors had done so, over time.
c) Mean Tie Ratio (MTR) - Average density of connections among the first n adopters of the feature, n = 1 to 100.
d) Number of adopters of the feature over time.
e) Average fraction of connections between adopter friends over time.

2) Regional Contagion Monitor: For nominated hashtags, we calculated the same measures used for the SCM. We base the CDF$_k$ and MTR measures on the friend graph as it existed at the beginning of the given month, using the `friends/ids` query time stored with the edges. For each hashtag, the analysis period began at the start of the first day the hashtag had more than ten tweets, and ended on midnight of the trending day. We compute these measures for the current day.

In addition, we also looked at how users that adopted a hashtag were distributed across the network. We used the Louvain community detection algorithm [21] to define communities within the friend network. For each hashtag, we then computed the *community entropy* as the entropy of the counts of adopting users in each community. Lower entropy is associated with the adopting users being restricted to fewer communities; higher entropy associated with users being spread out across more communities. We called this the hashtag's community entropy.

F. Reporting

The contagion monitor generates a report of all metrics for each feature and also saves plots of them as PDFs.

IV. Results

A. Framework for Evaluating the Contagion Monitors

Here we frame our evaluation under two key science questions. CDF$_k$, MTR, and entropy relate to first, R1, and the remaining two relate to second.

R1. Do high-threshold social contagions spread through social networks in a fundamentally different way than low-threshold social contagions?
R2. Do high-threshold social contagions "go viral" and begin to spread really quickly through social networks in certain conditions?

For R1, the contagion monitors provide a formal encoding of how hashtags spread through the network, via metrics a) and c) in section III-E1. But they do not provide a human label for what a "high-threshold social contagion" is. For such human labels, we turn to previous research (section II-B), which suggests that political and social movements tend to spread like high-threshold contagions [5], [7]. This observation naturally suggests that we can evaluate the contagion monitors by generating human labels for whether a hashtag represents a political or social movement, and so we do (section IV-B).

For R2, the contagion monitors provide a formal encoding of our theoretical assumptions about conditions when a contagion might "go viral," via metrics d) and e) in section III-E1. But they do not provide a human label for what a "viral" contagion is. We have performed preliminary investigations of human labels for contagion virality and found, not surprisingly, that human labelers do not successfully identify viral vs. non-viral contagions. As a specific test, we compared human labels of hashtags to the number of tweets using these hashtags and found no correlation. Therefore, R2 will require a separate evaluation, which we leave to future work.

B. Mechanical turk annotation

We used Amazon Mechanical Turk (AMT) to obtain human annotations of the hashtags nominated by the SCM and the RCM. AMT allows researchers to post Human Intelligence Tasks (HITs), asking AMT workers (turkers) to perform a task for a small payment. It is a common tool for obtaining labeled data at scale, in our case, to assess labels for adoption threshold. For our HITs, turkers were asked to read a set of tweets corresponding to the date a hashtag was nominated, answer a number of questions about the topic (or topics) associated with the hashtag, and give their judgements about the use of the hashtag by Twitter users.

For the SCM, we restricted our analysis to hashtags that were associated with English language tweets since many turkers are English speakers and we wanted to avoid having to identify turkers fluent in other languages. Language identification was carried out by running a language identification tool[4] on pseudo documents created by concatenating all of the tokenized tweets returned for a hashtag (Public Twitter API) after removing sentence punctuation, URLs, and emojis. We used language predictions for documents with 25 or more tokens and found 866 hashtags identified as English. For the RCM, we did no screening of language since most Twitter content we have observed from Nigeria has been in English or Nigerian Pidgin English.

For the SCM hashtags, we split the hashtags into equal-sized sets based on their likelihood of being a complex contagion. We created a "complex contagion score" (CCS) as follows:

$$\text{CCS} = \begin{cases} 0 & log_{10}(\bar{\rho}_{\text{ first 100 adopters}}) \leq -3 \cap CDF_k(k \geq 2) \geq 0.7 \\ 1 & log_{10}(\bar{\rho}_{\text{ first 100 adopters}}) > -3 \cup CDF_k(k \geq 2) < 0.7 \\ 2 & log_{10}(\bar{\rho}_{\text{ first 100 adopters}}) > -3 \cap CDF_k(k \geq 2) < 0.7 \end{cases} \tag{1}$$

The score is an integer that ranges from 0 (unlikely to be a complex contagion) to 2 (very likely to be a complex contagion). The set size was set to the number of hashtags that had a complex contagion score of 2. Hashtags with a score of two were the most interesting to us since they had a high observed adoption threshold and a spread within a dense initial network of users, both of these measures being theoretical indicators of complex contagion. There were 127 with a complex contagion of 2. Since we were using three hashtags per HIT, the total number of hashtags of each set was 126 giving us 126 HITs for 378 hashtags. For the RCM, 648 were selected at random without regard to contagion scores.

We deployed the HITs and required nine assignments per HIT, meaning that we asked for judgments from nine different turkers. There is no convenient way to restrict how many HITs a turker does for a particular batch of HITs. This has consequences since a few prolific turkers could be responsible for a disproportionate number of annotations. Using qualifications (a means provided by AMT to control turker's access to HITs) we implemented an approach using JavaScript and a custom RESTful interface that allowed us to disqualify a turker after they did a certain number of HITs. In this case, we restricted turkers to one SCM HIT and ten RCM hashtags.

Each HIT contained three hashtags, with each hashtag appearing on a separate page. At the top of each page there was a link to a Twitter Advanced Search query showing tweets containing the hashtag for the analysis period. Turkers were required to click on this link, which opened in a separate tab, and read a sample of the tweets using the hashtag before answering any of the questions. They were also required to answer all of the questions (shown in Table I) before advancing to the next page or submitting the HIT. We paid $1.80 per HIT (based on trial runs, a single HIT took about 10 minutes giving turkers an estimated hourly rate of $10.80). Each HIT allowed for nine assignments to unique turkers.

We ran a total of 640 HITs and obtained annotations from 1,084 turkers across both sets of hashtags. A pair of check questions determined whether turkers were paying sufficient attention to the task. Results from assignments where workers failed this check were not used. Overall, the average number of assignments used from each HIT (out of a possible nine assignments) was 8.20 (s.d. = 1.234) for the RCM hashtags and 7.89 (s.d. = 1.457) for the SCM ones. The average agreement (the maximum number of responses to a question - casting all questions as a binary choice - that were in agreement) was 6.83 (s.d. = 1.672) and 6.36 (s.d. = 1.692), respectively.

C. *Analysis of annotated hashtags*

From our sets of annotated hashtags we investigated the relationship between the contagion measures ($CDF_k(k \geq 2)$, MTR, and, for RCM, the community entropy) and the question responses obtained from AMT. We sought to learn how different ways of indicating threshold (from the responses) mapped onto the metrics. So, we specifically looked at regression models using the responses as dependent variables and the measures as independent variables to evaluate model fit and the significance of the measures. We then trained and evaluated classifiers for predicting positive question responses (as binary labels) using the measures as features.

1) Spam Filtering: While the SCM candidate hashtag selection infrastructure filters out spam-related hashtags, the RCM does not - other than initially filtering out hashtags spread exclusively by extremely productive or isolated, friendless accounts. We took the extra step of analyzing bot activity in the RCM data after noting the presence of hashtags with very low $CDF_k(k \geq 2)$ values (a high percentage of instigators) and low community entropy values (the accounts are isolated to a relatively small number of communities). Further examination of this data found that multiple accounts were tweeting the same links and hashtags simultaneously[9]. A group of these accounts showed the same behavior over a number of hashtags and had a higher than expected connectivity within the friend network, suggesting they were part of a bot network. Forty-nine of the 648 hashtags were almost completely dominated by these accounts. They were removed, giving us a set of 599 hashtags for the following analysis, together with the hashtags from the SCM. The bot activity in the RCM dataset is of interest in and of itself, but will be left to follow on research.

2) Linear Regression: We calculated the mean of the nine responses to each HIT question for both sets of hashtags. For the yes/no questions, we interpreted a mean value of 0.5 or more as a positive response; for the five-level Likert scale questions, a mean value of 2.5 or more was taken as positive. Questions with less than 5% positive responses were ignored.

Linear regression models were generated for each question using all three measures. For brevity, we focus on four

[4] https://github.com/saffsd/langid.py

[9] Millisecond resolution is not available in the metadata returned by the `search` API, so 'simultaneously,' in this case, is at resolution of one second.

TABLE I: HIT questions.

1. After reading the tweets, please provide a careful description of what the hashtag is about. (text response)
2. What general topics does this hashtag reference? Check all that apply. (Multiple choice[5] + optional text response)
3. Have you heard of this hashtag outside of this HIT? (Yes/No)
4. How often have you seen this hashtag used in a Tweet or Retweet? (not counting the tweets you read for this task) (five-point Likert scale: strongly disagree - strongly agree)
5. Does this hashtag get used in tweets from two or more groups with opposing messages? (Yes/No)
6. The hashtag is controversial (defined as "prolonged public disagreement or heated discussion") (five-point Likert scale strongly disagree - strongly agree)
7. The hashtag expresses an opinion. (five-point Likert scale strongly disagree - strongly agree)
8. The hashtag expresses an opinion that matches my own. (six-point Likert scale strongly disagree - strongly agree + no opinion expressed)
9. I find this hashtag funny. (five-point Likert scale strongly disagree - strongly agree)
10. I find this hashtag interesting. (five-point Likert scale strongly disagree - strongly agree)
11. I would be uncomfortable using this hashtag. (five-point Likert scale strongly disagree - strongly agree)
12. I would be uncomfortable if my friends used this hashtag. (five-point Likert scale strongly disagree - strongly agree)
13. I find this hashtag to be offensive. (five-point Likert scale strongly disagree - strongly agree)
14. Using this hashtag would send a disturbing message. (five-point Likert scale strongly disagree - strongly agree)
15. This hashtag is related to a political movement. (five-point Likert scale strongly disagree - strongly agree)
16. Many Twitter users would be concerned about offending other users if they used this hashtag. (five-point Likert scale strongly disagree - strongly agree)
17. Check all of the following reasons why someone might NOT want to use this hashtag. (Multiple choice[6] + optional text response)
18. How contagious is this hashtag? (defined as spreading from one person to another) (four-point Likert scale[7])
19. My friends would start using this hashtag… (choose the earliest option that is true). (six-point Likert scale[8])

responses: controversial and political (which should correspond to high-threshold contagions) and news events (labeled as "events") and sports (which should correspond to low-threshold contagions). The results are in Tables II and III.

The results show that $CDF_k(k \geq 2)$ has a positive and significant relationship with controversial and political hashtags, and a negative and significant relationship with sports and news events hashtags (except for the RCM monitor, where the measure has a positive and significant relationship with news events hashtags).

For both the monitors, the coefficients for MTR are negative and statistically significant. This result is in contrast with complex contagion theory, which indicates that political and controversial hashtags would be more likely to emerge in denser network neighborhoods. We investigated this pattern further by visualizing the relationship between MTR and the hashtag labels. It follows the general pattern shown in Fig. 1.

The pattern shows an increase in Controversial Score for $0 \leq MTR \leq 0.1$ and a rapid decrease in Controversial score thereafter. We manually examined hashtags identified by the RCM with $MTR > 0.1$. A sample is shown in Table IV. Many of these hashtags seem related to marketing or spam. While Fig. 1 does not suggest an explicit cutoff for MTR, these observations suggest that the cause of the negative coefficient for MTR in tables II, III is a set of spam-related hashtags with extremely high mean tie ratios.

Finally, the results show that Entropy does not have a statistically significant relationship with any of the labels, except for controversial, where the relationship is negative.

3) Classification: We constructed a classifier for HIT questions, using $CDF_k(k \geq 2)$, MTR, and Community Entropy (where applicable). For the classification experiments we used

TABLE II: Regression results for SCM hashtags

Question	Intercept	1-CDF(k=2)	Mean Tie Ratio	Adjusted R-squared
controversial	1.77			0.00
	1.06	3.00***		0.22
	1.12	3.38***	-7.81***	0.25
news events	0.36			0.00
	0.37	-0.05(0.55)		0.00
	0.38	0.03(0.74)	-1.54**	0.03
politics	0.42			0.00
	0.18	1.01***		0.23
	0.21	1.18***	-3.41***	0.29
sports	0.09			0.00
	0.17	-0.33***		0.05
	0.18	-0.31***	-0.43(0.41)	0.05

TABLE III: Regression results for RCM hashtags

Question	Intercept	1-CDF(k=2)	Mean Tie Ratio	Entropy	Adjusted R-squared
controversial	1.05				0.00
	0.91	0.55***			0.03
	0.91	0.94***	-1.62***		0.08
	1.45	0.83***	-2.15***	-0.39**	0.10
news events	0.24				0.00
	0.19	0.18***			0.03
	0.19	0.29***	-0.47***		0.08
	0.11	0.31***	-0.39***	0.06(0.13)	0.10
politics	0.12				0.00
	0.02	0.35***			0.03
	0.03	0.44***	-0.37***		0.08
	0.12	0.42***	-0.46***	-0.07(0.06)	0.10
sports	0.32				0.00
	0.49	-0.61***			0.03
	0.49	-0.41***	-0.84***		0.08
	0.32	-0.38***	-0.68***	0.12(0.08)	0.10

TABLE IV: Sample hashtags with high MTR

streetmediapromotions, formularbyweflo, formulaoutsoon, doit, goosebumpsvideobytobe, watchformulavideo, akwaibomtotheworld, formulavideo, lilayunchangeable, nozippy, pmfa2017, viktohybnl, smirnoffnightuyo, kissmebyoludre, sirehabbiibbpr, nammkpohmfo, thecypher, ipoetry, factswithkulqee, 2daystorepurclubuyo, afrima2017, smoothsummersplash, marryjuanabykamartachio, hypaft9ice

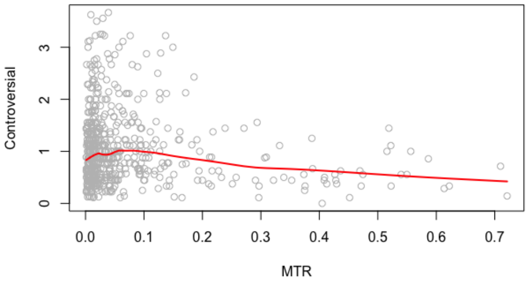

Fig. 1: Mean Tie Ratio vs. Controversial Score for RCM.

TABLE V: Classification results for SCM and RCM hashtags.

Question	Model	SCM				RCM			
		Sensitivity	Specificity	Accuracy	Baseline Accuracy	Sensitivity	Specificity	Accuracy	Baseline Accuracy
controversial	1-CDF(k=2)	0.63	0.78	**0.72**	0.63	0.45	0.68	0.67	0.95
	1-CDF(k=2) + MTR	0.62	0.76	**0.71**	0.63	0.73	0.68	0.68	0.95
	1-CDF(k=2) + MTR + ENTROPY					0.68	0.70	0.70	0.95
events	1-CDF(k=2)	0.74	0.37	0.48	0.71	0.44	0.72	0.68	0.87
	1-CDF(k=2) + MTR	0.57	0.68	0.87	0.87	0.57	0.68	0.68	0.87
	1-CDF(k=2) + MTR + ENTROPY					0.62	0.62	0.62	0.87
sports	1-CDF(k=2)	0.81	0.52	0.54	0.93	0.45	0.46	0.57	0.68
	1-CDF(k=2) + MTR	0.82	0.51	0.53	0.93	0.73	0.49	0.61	0.68
	1-CDF(k=2) + MTR + ENTROPY					0.68	0.55	0.66	0.68
politics	1-CDF(k=2)	0.61	0.76	**0.69**	0.53	0.65	0.75	0.74	0.90
	1-CDF(k=2) + MTR	0.61	0.76	**0.69**	0.53	0.66	0.74	0.73	0.90
	1-CDF(k=2) + MTR + ENTROPY					0.68	0.68	0.74	0.90

a 70%–30% training–test split and trained a model using a support vector machine with a linear kernel. The positive classes (turkers said ”yes” or agreed with the the question) were the minority class for all questions. The average positive response rate was 0.28 (0.137) for the SCM hashtags and 0.16 (0.114) for the RCM. To address this class imbalance we used minority class oversampling using the SMOTE algorithm [22]. We then ran the model against the held-out test set and calculated sensitivity (the true positive rate or recall), the sensitivity (the true negative rate), the accuracy, and the accuracy of a baseline classifier that always selects the majority label. This procedure was repeated on 10 random training–test splits and the performance measures were averaged across trials. For each trial we used a grid search to find the best value of C, the regularization term for the linear kernel. The results for a selection of the questions from the the two sets are shown in Table V. The table shows that for the SCM, $1-CDF(k=2)$ produces the best results by accuracy, though for the RCM, the inclusion of MTR as a feature provides an accuracy boost. Overall, both classifiers show best performance at predicting controversial and politics-related hashtag labels.

V. Discussion and Conclusion

The regional (RCM) and streaming (SCM) implementations of the social contagion monitor that we have built successfully process a large volume of data to identify emerging hashtags, representative of low- and high-threshold contagions. These tools let us test complex social theories at a large scale and over long periods of time. We hope to extend the monitors further to collect other public data and integrate additional social theories, such as homophily, into their analytic toolkit.

Our evaluation of the contagion monitors shows that hashtags with more social reinforcement are more likely to be labeled as political and controversial. The streaming contagion monitor also shows that hashtags with less social reinforcement are more likely to be labeled as news events and sports. These findings are in precise accordance with complex contagion theory [7]. The regional contagion monitor diverges from this pattern mildly to show that hashtags with more social reinforcement are also more likely to be labeled as news events. It is interesting to consider whether sharing news events is a higher-threshold behavior in Nigeria than globally, or whether a different confound accounts for this pattern.

We have observed a surprising negative relationship between Mean Tie Ratio and controversial and political hashtags. This relationship is inconsistent with complex contagion theory, which states that complex contagions are more likely to arise in more dense neighborhoods. However, a closer investigation of the relationship shows that it is driven by hashtags with extremely high MTR, which may be spam related. We do not see this finding as ultimately at odds with complex contagion theory, but an interesting development of the same. Perhaps organizers of spam related hashtags form dense networks to facilitate the spread of content due to social reinforcement.

Our classification results demonstrate that a) both monitors perform best when predicting labels of controversial and political hashtags, and b) most accuracy gain comes from including $1-CDF(k=2)$ as a feature, with MTR and Entropy contributing far less to accuracy gain. These results are also consistent with our interpretation of complex contagion theory: we use CDF_k as a proxy measure for contagion threshold, so it is the key differentiator between complex and simple contagions. The other two metrics provide circumstantial evidence for contagion threshold, as complex contagions can leverage redundant ties more easily in adopter networks with higher MTR and lower community entropy. Since we only evaluated linear classifiers, it is possible that higher-dimensional classifiers may achieve more accuracy gain from these measures. We leave this investigation for future work.

The classification performance of the streaming contagion monitor beats the baseline for controversial and political hashtags using only a linear classifier and two or three features. To our knowledge, our findings have never been replicated with the same instrument across multiple social, cultural, and linguistic settings and our paper is also the first to label hashtag categories and evaluate complex contagion theory in the context of a productionalized tool.

Overall, we have found the social contagion monitors to successfully identify emerging low- and high-threshold movements in both regional and global Twitter settings. One shortcoming of our approach, beyond the inherent limitations of using ”digital traces” to model diffusion in social networks, is that we model complex contagions as separate events. However, movements that lead to behavioral change can inspire one another, requiring them to be modeled together and not in isolation. We leave such modeling efforts for future study.

We recognize further that this tool may surface dis- or mis-information instead of real transformative movements, resulting from our focus on data quantity over data quality. We hope in future work, to combine the contagion monitor with methods for verifying the quality of an information stream, either in automation [23] or along with expert human review.

A sign of the Contagion Monitor's continuing predictive ability was seen in the productionalized version's use in 2018 to discover two rising artists go from fringe to mainstream. We had added the ability to study Twitter @mentions to the monitor. In March of 2018 we analyzed social contagions in a commercial environment – Twitter activity around the South By Southwest (SXSW) music festival, to look for up and coming musicians and artists (where supporting a new artist has a higher social cost due to their lower popularity). We found the Twitter handles of Desus and Mero, a comedian and a DJ, had both a high adoption threshold and had reached critical mass. We followed these artists over a longer time period, and observed how their rapid spread on social media was followed by commercial success [24], including a successful performance tour and a dedicated program on the Showtime channel. The performance of the contagion monitor suggests that it can be a powerful tool for identifying both politically and commercially relevant content in social media.

Appendix

See Tables VII and VI for topics and Likert scale descriptions for questions listed in Table I.

References

[1] V. Barash, C. Cameron, and M. Macy, "Critical phenomena in complex contagions," *Social Networks*, vol. 34, no. 4, pp. 451 – 461, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873312000111

[2] C. Fink, A. Schmidt, V. Barash, C. Cameron, and M. Macy, "Complex contagions and the diffusion of popular twitter hashtags in nigeria," *Social Networks Analysis and Mining*, vol. 6, no. 1, 2016.

[3] J. Cheng, L. Adamic, P. A. Dow, J. Kleinber, and J. Leskovec, "Can cascades be predicted," in *Proc. 23rd International World Wide Web Conference*, 2014.

[4] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "The dynamics of protest recruitment through an online network," *Scientific Reports*, vol. 1, no. 197, 2011.

[5] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011, pp. 695–704.

[6] V. A. Traag, "Complex contagion of campaign donations," *PLoS ONE*, vol. 11, no. 4, 2016.

[7] D. Centola and M. Macy, "Complex contagions and the weakness of long ties1," *American Journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.

[8] S. Morris, "Contagion," *Review of Economic Studies*, vol. 67, pp. 57–78, 2000.

[9] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[10] K. Nahon and J. Hemsley, *Going viral*. Polity, 2013.

[11] I. University. (2017) Observatory on social media. [Online]. Available: http://truthy.indiana.edu/

[12] S. University. (2017) Nifty. [Online]. Available: http://snap.stanford.edu/nifty/

[13] C. Hexagon. (2017) Crimson hexagon. [Online]. Available: http://www.crimsonhexagon.com

[14] Graphika. (2017) Graphika. [Online]. Available: http://www.graphika.com

[15] Salesforce. (2017) Social studio by salesforce. [Online]. Available: https://www.salesforce.com/products/marketing-cloud/channels/social-media-marketing/

[16] M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. (2013) Teens, social media, and privacy. pew research center internet, science and tech. [Online]. Available: http://www.pewinternet.org/2013/05/21/teens-social-media-and-privacy/\#fn-67-1

[17] Health and H. Services. (2017) Office for human research protections. subpart d. additional protections for children involved as subjects in research. [Online]. Available: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html\#subpartd

[18] H. R. Watch. (2016) "tell me where i can be safe" the impact of nigeria's same sex marriage (prohibition) act. [Online]. Available: https://www.hrw.org/report/2016/10/20/tell-me-where-i-can-be-safe/impact-nigerias-same-sex-marriage-prohibition-act

[19] C. Fink, A. Schmidt, V. Barash, J. Kelly, C. Cameron, and M. Macy, "Investigating the observability of complex contagion in empirical social networks." in *In Proceedings of the 10th International Conference on Weblogs and Social Media*, 2016.

[20] J. Kelly, V. Barash, K. Alexanyan, B. Etling, R. Faris, U. Gasser, and J. Palfrey, "Mapping russian twitter," *Berkman Center Research Publication*, no. 3, 2012.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008. [Online]. Available: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[23] C. François, V. Barash, and J. Kelly, "Measuring coordinated vs.spontaneous activity in online social movements," 2017, preprint available on SocArxiv - https://osf.io/aj9yz/.

[24] Graphika. (2019) https://graphika.com/posts/the-desus-and-mero-story/. [Online]. Available: https://graphika.com/posts/the-desus-and-mero-story/

TABLE VI: Question Topics from Table I

Question	Topics
Q–2	Movie; TV; News Event; Music; Celebrity; Sport; Health; Religion; Marketing Campaign; Hacking or Cyberattack; Politic; Divisive Social or Moral Issue; Sexuality; Corruption; Injustice; Crime/Police; War or Armed Conflict; Protest or Rally; Criticism of Society; Criticism of Government; Meme; Joke
Q–17	Offensive; Disturbing; Polarizing; Using this hashtag could hurt one's reputation; Unfamiliar; Uninteresting; The hashtag uses inappropriate language; The hashtag is used by people that one might not want to be associated with.

TABLE VII: Likert Scales for Questions in Table I

Question	Scale Descriptions
Q–18	1. Not contagious at all — People would not use this hashtag no matter how many others were using it 2. Slightly contagious — People would only use this hashtag if they saw many people using it 3. Moderately contagious — People might decide to use this hashtag if they noticed a few people were using it 4. Highly contagious — People would use this hashtag as soon as they saw someone use it for the first time
Q–19	1. If they heard about it from any random source 2. If one of their friends was using it 3. If a few of their friends were using it 4. If many of their friends all were using it 5. If nearly all of their friends were using it 6. Under no circumstances