Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter

Thayer Alshaabi,[1,2,∗] Jane L. Adams,[1,2,†] Michael V. Arnold,[1,2,†] Joshua R. Minot,[1,2,†] David R. Dewhurst,[1,2,3] Andrew J. Reagan,[3] Christopher M. Danforth,[1,2,4] and Peter Sheridan Dodds[1,2,4,‡]

[1] *Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405.*
[2] *Computational Story Lab, University of Vermont, Burlington, VT 05405.*
[3] *MassMutual Data Science, Boston, MA 02110*
[4] *Department of Mathematics & Statistics, The University of Vermont, Burlington, VT 05405.*
(Dated: July 28, 2020)

In real-time, Twitter strongly imprints world events, popular culture, and the day-to-day; Twitter records an ever growing compendium of language use and change; and Twitter has been shown to enable certain kinds of prediction. Vitally, and absent from many standard corpora such as books and news archives, Twitter also encodes popularity and spreading through retweets. Here, we describe Storywrangler, an ongoing, day-scale curation of over 100 billion tweets containing around 1 trillion 1-grams from 2008 to 2020. For each day, we break tweets into 1-, 2-, and 3-grams across 150+ languages, record usage frequencies, and generate Zipf distributions. We make the data set available through an interactive time series viewer, and as downloadable time series and daily distributions. We showcase a few examples of the many possible avenues of study we aim to enable including how social amplification can be visualized through 'contagiograms'.

I. INTRODUCTION

Our collective memory lies in our recordings—in our written texts, artworks, photographs, audio, and video—and in our retellings and reinterpretations of that which becomes history.

The relatively recent digitization of historical texts, from books [1–4] to news [5–7] to folklore [8–11] to governmental records [12], has enabled compelling computational analyses across many fields [9, 13, 14]. But the large-scale constructions of historical corpora often entirely fail to encode a fundamental characteristic: Popularity. How many people have read a text? How many have retold a news story to others?

For text-based corpora, we are confronted with the challenge of sorting through three levels of popularity of n-grams—sequences of n "words" in a text that are formed by contiguous characters, numerals, symbols, emojis, etc.

The first level of popularity is dictated by whether or not an n-gram is simply part of a text's lexicon—the level of types [15]. The vocabulary of a text gives a base sense of what that text may span meaning-wise.

The second level of popularity is that of recorded tokens in a corpus of texts unindexed by fame—the realizations of types [15]. For texts, it has long been well known that n-gram frequency-of-usage (or Zipf) distributions are heavy-tailed [16]. Problematically, this essential character of natural language is readily misinterpreted as indicating cultural popularity. For a prominent example, the Google Books n-gram corpus [1], which in part pro-

vides inspiration for our work here, presents year-scale, n-gram frequency time series where each book, in principle, counts only once [2]. All cultural fame is stripped away. The words of George Orwell's 1984 or Rick Riordan's Percy Jackson books, indisputably read and re-read by many people around the world, count as equal to the words in the least read books published in the same years. And yet, time series provided by the Google Books n-gram viewer have regularly been erroneously conflated with the changing interests of readers (e.g., the apparent decline of sacred words [17]; see also [2, 18–21]). Further compounded with an increase of scientific literature throughout the 20th Century, the corpus remains a deeply problematic database for investigations of sociolinguistic and cultural trends.

The third, most important, and most difficult to measure level is that of cultural popularity. For a given book, we would want to know sales of the book over time, how many times the book has been actually read, and to what degree a book becomes part of broader culture.

Large-scale corpora capturing all three levels of popularity exist [14] but in general are hard to compile as the relevant data is either prohibitively expensive or closed (e.g., Facebook), and, even when accessible, may not be consistently recorded over time (e.g., Billboard's Hot 100).

Now well into the age of the internet, our recordings are vast, inherently digital, and capable of being created and shared in the moment. People, news media, governmental bodies, corporations, bots, and many other entities all contribute constantly to giant social media platforms. When open, these services provide an opportunity for us to attempt to track myriad statements, reactions, and stories of large populations in real time.

And crucially, when sharing and commenting mechanisms are native to a social media platform, we can quantify social amplification and thereby measure all three

∗ thayer.alshaabi@uvm.edu
† Equal contribution
‡ peter.dodds@uvm.edu

levels of popularity.

For a social media source, we advocate for the use of Twitter qua text for a number of reasons, while acknowledging its limitations.

First, Twitter acts as a distributed sociotechnical sensor system [22–24]. People can record and share events anchored by one fundamental piece of metadata: the time stamp. The observer base for major events is now no longer limited to those physically present because of growing, decentralized live-streaming through various social media platforms.

Second, Twitter has become an essential outlet for news sources of all kinds, including mistaken reporting and those deliberately spreading misinformation. With complex, ever-expanding fabric of time-stamped messages, Twitter allows news storylines to be surfaced and traced in real time. All major news events and stories—from natural disasters [25–29] to political events [30]—are represented on Twitter.

Third, aside from news, Twitter is suffused with discussions of sports, popular culture (e.g., K-pop and Star Wars), sports, and the quotidian. Twitter carries both the voices of famous individuals—political figures and celebrities—and the expressions of the many.

Fourth, while we are limited to the recent past when using social media, the temporal resolution for Twitter posts is at the scale of milliseconds, far finer than would be reasonably needed to explore sociocultural phenomena or reconstruct major events.

Fifth, Twitter enables and records social interactions through replies, favoriting, and retweets. For n-grams, we capitalize on Twitter presenting us with social amplification through retweets and quote tweets. For each day and across languages, we create Zipf distributions for: (1) n-grams from organic tweets (OT), excluding all retweeted material (RT); and (2) n-grams from all Twitter messages (AT). For each day, we then have the three key levels of popularity: n-gram lexicon, n-gram usage in organic tweets, and n-gram in tweets with social amplification. Our data set allows researchers to explore transitions in social amplification by reconstructing n-gram Zipf distributions with a tunable fraction of retweets.

Finally, like other major social media platforms, Twitter's collective voice is important in and of itself: What Twitter algorithmically curates and presents to users evidently matters, and news outlets routinely report on what is trending on Twitter.

Like any large-scale text corpus, Twitter has substantive limitations, some of which are specific to Twitter itself. Twitter's user base, while broad, is not perfectly representative of a populace [31], is moreover compounded by the mixing of voices from people, organizations, and bots, and has evolved over time as new kinds of users have joined.

Increasingly by design, geographic information is limited on Twitter as are user demographics, though some aspects may be gleaned indirectly [32–35]. Regardless, in our first curation of Twitter n-grams, we purposeful-ly do not attempt to incorporate any metadata beyond identified language into our n-gram database.

We structure our paper as follows. In Sec. II, we describe in brief our data set and the Storywrangler site for Twitter which provides day-scale n-gram time series data sets for n=1, 2, 3. both as time series and as daily Zipf distributions. In Sec. III, we showcase three groups of example analyses, arranged by increasing complication: Simple n-gram rank time series (Sec. III A); Contagiograms, time series showing social amplification (Sec. III B); and an example set of case studies bridging n-gram time series with disparate data sources to study famous individuals, box office success, and social unrest (Sec. III C). In our concluding remarks in Sec. IV, we outline some potential future developments for Storywrangler.

II. OVERVIEW OF STORYWRANGLER DATA SET AND INTERACTIVE SITE

A. n-gram data set

We draw on a storehouse of messages comprising roughly 10% of all tweets collected from 2008/09/09 onwards, and covering around 150+ languages (see Data and Methods, Sec. S1). In previous work [36], we described how we re-identified the languages of all tweets in our collection using FastText [37], uncovering a general increase in retweeting across Twitter over time. A uniform language re-identification was needed as Twitter's own real-time identification algorithm was introduced in late 2012 and then adjusted over time, resulting in temporal inconsistencies for long-term streaming collection of tweets [38].

For each Coordinated Universal Time (UTC) day t and for each language ℓ, we break tweets into n-grams with n=1, 2, and 3. We accommodate all unicode characters including emojis, contending with punctuation as fully as possible (see [36] for details). Date and language are the only metadata we incorporate into our database. For user privacy in particular, we discard all other information associated with a tweet.

We derive three essential measures for each n-gram: raw frequency (or count), normalized frequency (interpretable as probability), and rank, generating the corresponding Zipf distributions [16]. We perform this process for all tweets (AT), organic tweets (OT), and (implicitly) retweets (RT). We then record n-grams along with ranks, raw frequencies, normalized frequencies for all tweets and organic tweets in a single file, with the default ordering according to n-gram prevalence in all tweets.

B. Notation and Measures

We write an n-gram by τ and a day's lexicon for language ℓ—the set of distinct n-grams found in all tweets
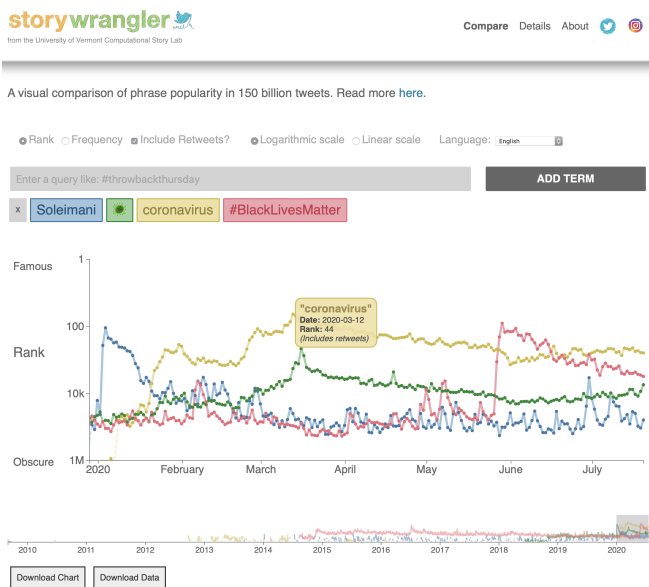
FIG. 1. Screenshot of the Storywrangler site showing example Twitter n-gram time series for the first half of 2020. The series reflect three global events: The assassination of Iranian general Qasem Soleimani by the United States on 2020/01/03, the COVID-19 pandemic (the virus emoji 🦠 and 'coronavirus'), and the Black Lives Matter protests following the murder of George Floyd by Minneapolis police ('#Black-LivesMatter'). The n-gram Storywrangler dataset for Twitter records the full ecology of text elements, including punctuation, hashtags, handles, and emojis. The default view is for n-gram (Zipfian) rank at the (GCT) day scale, a logarithmic scale, and for retweets to be included; these settings can be respectively switched to normalized frequency, linear scale, and organic tweets (OT) only. The displayed time range can be adjusted with the selector at the bottom, and all data is downloadable.

(AT) for a given date t—by $\mathcal{D}_{t,\ell;n}$. We write n-gram raw frequency as $f_{\tau,t,\ell}$, and compute its usage rate in all tweets written in language ℓ as

$$p_{\tau,t,\ell} = \frac{f_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}_{t,\ell;n}} f_{\tau',t,\ell}}. \tag{1}$$

We further define the set of unique language ℓ n-grams found in organic tweets as $\mathcal{D}_{t,\ell;n}^{(\mathrm{OT})}$, and the set of unique n-grams found in retweets as $\mathcal{D}_{t,\ell;n}^{(\mathrm{RT})}$ (hence $\mathcal{D}_{t,\ell;n} = \mathcal{D}_{t,\ell;n}^{(\mathrm{OT})} \cup \mathcal{D}_{t,\ell;n}^{(\mathrm{RT})}$). The corresponding normalized frequencies for these two subsets of n-grams are then:

$$p_{\tau,t,\ell}^{(\mathrm{OT})} = \frac{f_{\tau,t,\ell}^{(\mathrm{OT})}}{\sum_{\tau' \in \mathcal{D}_{t,\ell;n}^{(\mathrm{OT})}} f_{\tau',t,\ell}^{(\mathrm{OT})}} \text{ and } p_{\tau,t,\ell}^{(\mathrm{RT})} = \frac{f_{\tau,t,\ell}^{(\mathrm{RT})}}{\sum_{\tau' \in \mathcal{D}_{t,\ell;n}^{(\mathrm{RT})}} f_{\tau',t,\ell}^{(\mathrm{RT})}}. \tag{2}$$

We rank n-grams by raw frequency of usage using fractional ranks for ties. The corresponding notation is:

$$r_{\tau,t,\ell}, \ r_{\tau,t,\ell}^{(\mathrm{OT})}, \text{ and } r_{\tau,t,\ell}^{(\mathrm{RT})}. \tag{3}$$

C. Storywrangler site

We make interactive times series based on our n-gram data set viewable at storywrangling.org. In Fig. 1, we show a screenshot of the site displaying rank time series for the first half of 2020 for 'Soleimani', the virus emoji 🦠, 'coronavirus', and '#BlackLivesMatter'. Ranks and normalized frequencies for n-grams are relative to n-grams with the same n, and we show time series on separate axes below the main comparison plot.

For each time series, hovering over any data point will pop up an information box, like the example shown in Fig. 1 for 'coronavirus' on 2020-03-12. Clicking on a data point will take the user to Twitter's search results for the n-gram for the span of three days centered on the given date.

All time series are shareable and downloadable through the site, as are daily Zipf distributions for the top 10^6 n-grams in each language. Retweets may be included (the default) or excluded, and the language, vertical scale, and time frame may all be selected.

III. EXAMPLE EXPLORATIONS

A. Basic rank time series

In Fig. 2, we show rank time series for eight sets of n-grams from all tweets (i.e., including retweets). The n-gram groups move from simple to increasingly complex in theme, span a number of languages, and display a wide range of sociotechnical dynamics. Because of an approximate obeyance of Zipf's law ($f \sim r^{-\theta}$) we observe that normalized frequency of usage time series match rank time series in basic form. We use rank as the default view for its straightforwardness.

Starting with time and calendars, Fig. 2A gives a sense of how years are mentioned on Twitter. The dynamics show an anticipatory growth, plateau, and then rapid decay, with each year's start and finish marked by a spike.

Figs. 2B and C show calendrically anchored rank time series for seasonal, religious, political, and sporting events that recur at the scale of years in various languages. Periodic signatures at the day, week, and year scale are prominent on Twitter, reflecting the dynamics of the Earth, moon, and sun. Easter (shown in Italian) in particular combines cycles of all three. Major sporting events produce time series with strong anticipation, and can reach great heights of attention as exemplified by a peak rank of $r = 3$ for 'Super Bowl' on 2014/02/02.

We move to scientific discoveries in Fig. 2D with the 2012 discovery of Higgs boson particle (blue), detection of gravitational waves (green), and the first imaging of a black hole (red). For innovations, we show the time series of '#AlphaGo'—the first artificial intelligence program to beat the human Go champion (orange), along with the development of CRISPR technology for editing
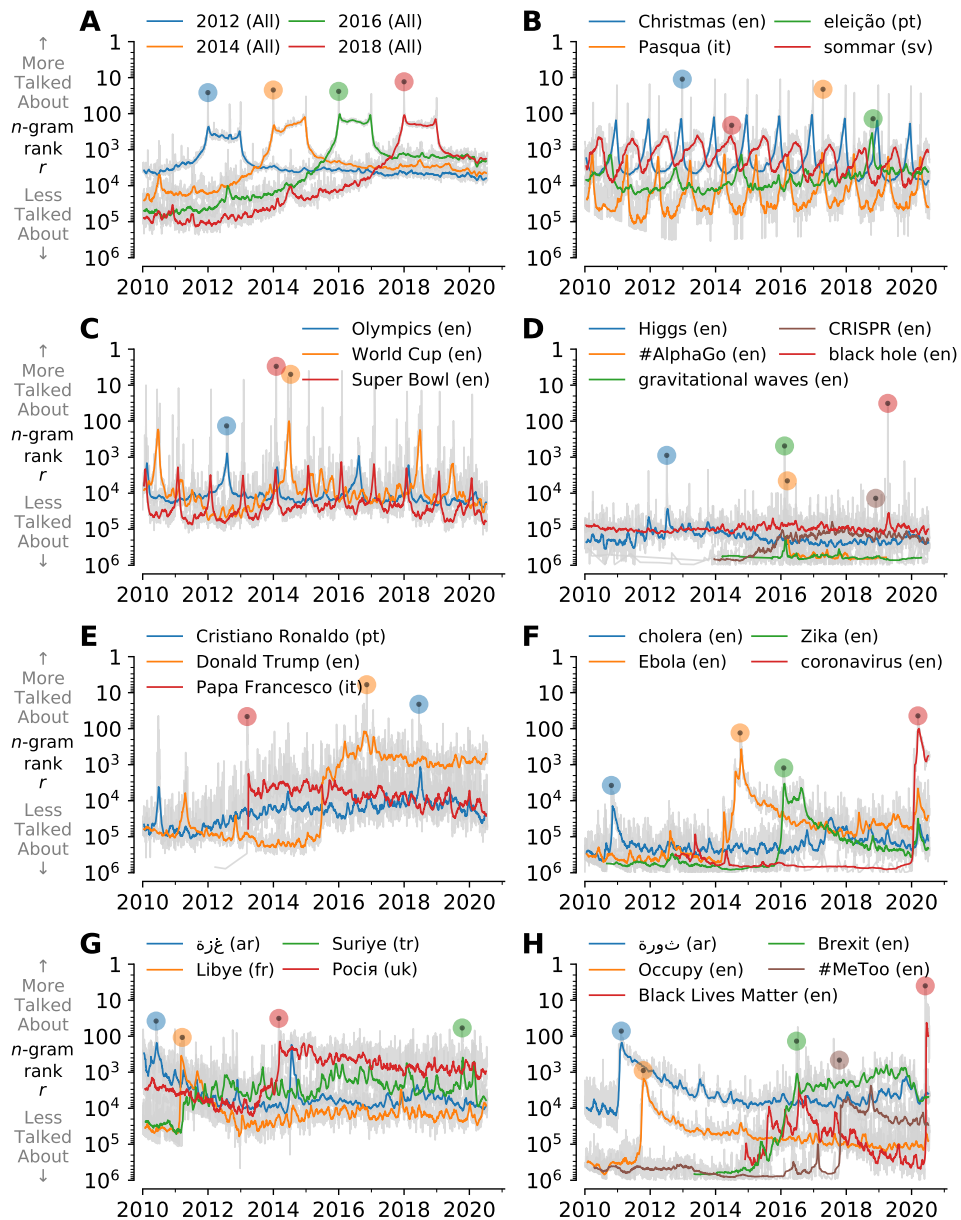
FIG. 2. **Thematically connected *n*-gram time series.** For each *n*-gram, we display daily rank in gray overlaid by a centered monthly rolling average (colored lines), and highlight the *n*-gram's overall highest rank with a solid disk. **A.** Anticipation and memory of calendar years for all of Twitter. **B.** Annual and periodic events: Christmas in English (blue), Easter in Italian (orange), election in Portuguese (green), and summer in Swedish (red). **C.** Attention around international sports in English: Olympics (blue), FIFA world cup (orange), and Super Bowl (red). **D.** Major scientific discoveries and technological innovations in English. **E.** Three famous individuals in relevant languages: Ronaldo (Portuguese), Trump (English), and Pope Francis (Italian). **F.** Major infectious disease outbreaks. **G.** Conflicts: Gaza in Arabic (blue), Libya in French (orange), Syria in Turkish (green), and Russia in Ukrainian (red). **H.** Protest and movements: Arab Spring (Arabic word for 'revolution', blue), Occupy movement (English, orange), Brexit campaign (English, green), #MeToo movement (English, brown), and Black Lives Matter protests (English, red).

genomes (brown). We see that time series for scientific advances generally show shock-like responses with little anticipation or memory. CRISPR is an exception for these few examples as through 2015, it moves to a higher, enduring state of being referenced.

Fame is the state of being talked about and famous individuals are well reflected on Twitter [39]. In Fig. 2E we show time series for the Portuguese football player Cristiano Ronaldo, the 45th US president Donald Trump, and Pope Francis (Papa Francesco in Italian). All three show enduring fame, following sudden rises for both Trump and Pope Francis. In the 2016 US election, 'Donald Trump' rose as high as rank $r = 6$ among all English 2-grams.

In Fig. 2F, we show example major infectious disease

outbreaks over the last decade. Time series for pandemics are shocks followed by long relaxations, resurging both when the disease returns in prevalance and also in the context of new pandemics. Cholera, ebola, and zika all saw elevation in discussion within the context of the COVID-19 pandemic.

In Fig. 2G, we show *n*-gram signals of regional unrest and fighting. The word for Gaza in Arabic tracks events of the the ongoing Israeli-Palestinian conflict. The time series for 'Libye' points to Opération Harmattan the 2011 French and NATO military intervention in Libya. Similarly, the time series for 'Syria' in Turkish indicates the dynamics of the ongoing Syrian civil war on the region, and the build up and intervention of of Russian military in Ukraine is mirrored by the use of the Ukrainian word

for 'Russia'.

In Fig. 2H, we highlight protests and movements. Both the time series for 'revolution' in Arabic and 'Occupy' in English show strong shocks followed by slow relaxations over the following years. The social justice movements represented by #MeToo and 'Black Lives Mattter' appear abruptly, and their time series show slow decays punctuated by shocks returning them to higher ranks. Black Lives Matter resurged after the murder of George Floyd, with a highest one day rank of $r = 4$ occurring on 2020/06/02. By contrast, the time series of 'Brexit', the portmanteau for the movement to withdraw the United Kingdom from the European Union, builds from around the start of 2015 to the referendum in 2016, and then continues to climb in the following years of complicated negotiations.

B. Contagiograms

While rank time series for n-grams give us the bare temporal threads that make up the tapestries of major stories, our data set offers more dimensions to explore. Per our introductory remarks on the limitations of text corpora, the most important enablement of our database is the ability to explore story amplification.

In Fig. 3, we present a set of six 'contagiograms'. With these expanded time series visualizations, we convey the degree to which an n-gram is retweeted both overall and relative to the background level of retweeting for a given language. We show both rates because retweet rates change strongly over time and variably so across languages [36].

Each contagiogram has three panels. The main panel at the bottom charts, as before, the rank time series for a given n-gram. For contagiograms running over a decade, we show rank time series in this main panel with month-scale smoothing (black line), and add a background shading in gray indicating the highest and lowest rank of each week.

The top two panels of each contagiogram capture the raw and relative social amplification for each n-gram.

First, the top panel displays the raw RT/OT balance, the monthly relative volumes of each n-gram in retweets (RT, orange) and organic tweets (OT, blue):

$$R_{\tau,t,\ell} = f_{\tau,t,\ell}^{(\mathrm{RT})} / \left(f_{\tau,t,\ell}^{(\mathrm{RT})} + f_{\tau,t,\ell}^{(\mathrm{OT})} \right). \qquad (4)$$

When the balance of appearances in retweets outweighs those in organic tweets, $R_{\tau,t,\ell} > 0.5$, we view the n-gram as nominally being amplified, and we add a solid background for emphasis.

Second, in the middle panel of each contagiogram, we display a heatmap of the values of the relative amplification rate for n-gram τ in language ℓ, $R_{\tau,t,\ell}^{\mathrm{rel}}$, over time. Building on from the RT/OT balance, we define $R_{\tau,t,\ell}^{\mathrm{rel}}$

as:

$$R_{\tau,t,\ell}^{\mathrm{rel}} = \frac{f_{\tau,t,\ell}^{(\mathrm{RT})} / \left(f_{\tau,t,\ell}^{(\mathrm{RT})} + f_{\tau,t,\ell}^{(\mathrm{OT})} \right)}{\sum_{\tau'} f_{\tau',t,\ell}^{(\mathrm{RT})} / \sum_{\tau'} \left(f_{\tau',t,\ell}^{(\mathrm{RT})} + f_{\tau',t,\ell}^{(\mathrm{OT})} \right)}, \qquad (5)$$

where the denominator gives the overall fraction of n-grams that are found in retweets on day t for language ℓ. For the initial version of contagiograms, we approximate the denominator by the fraction of messages that are retweets (we count new material appearing in quote tweets as separate organic tweets). While still averaging at month scales, we now do so based on day of the week. Shades of red indicate that the relative volume of n-gram τ is being socially amplified over the baseline of retweets in language ℓ, $R_{\tau,t,\ell}^{\mathrm{rel}} > 1$, while gray encodes the opposite, $R_{\tau,t,\ell}^{\mathrm{rel}} < 1$.

The contagiogram in Fig. 3A for the word for 'kevät', 'spring' in Finnish, shows an expected annual periodicity. The word has a general tendency to appear in organic tweets more than retweets. But this is true of Finnish words in general, and we see that from the middle panel that kevt is in fact relatively, if patchily, amplified when compared to all Finnish words. For the anticipatory periodic times series in Fig. 3B, we track references to the 'Carnival of Madeira' festival—held forty days before Easter in Brazil. We see 'Carnival' has become increasingly amplified over time, and has been relatively more amplified than Portuguese words except for 2015 and 2016.

By etymological definition, renowned individuals should feature strongly in retweets ('renown' derives from 'to name again'). Lionel Messi has been one of the most talked about sportspeople on Twitter over the last decade, and Fig. 3C shows his 2-gram is strongly retweeted, by both raw and relative measures. (See also Fig. S1F for the K-pop band BTS's extreme levels of social amplification.)

Some n-grams exhibit a consistent weekly ampflification signal. For example, '#TGIF' is organically tweeted on Thursdays and Fridays, but retweeted more often throughout the rest of the week (Fig. 3D). Increased advertising on those two days for the eponymous restaurant chain would offer one explanation.

Routinely, n-grams will take off in usage and amplification due to global events. In Fig. 3E, we see 'virus' in French tweets holding a stable rank throughout the 2010s before jumping in response to the COVID-19 pandemic, and showing mildly relatively increased amplification levels. The word 'Brexit' in German has been prevalent from 2016 on, balanced in terms of organic tweet and retweet appearances, and generally relatively more spread than German 1-grams.

The contagiograms in Fig. 3 give just a sample of the rich variety of social amplification patterns that appear on Twitter. We include some further examples in the supplementary material in Figs. S1 and S2.

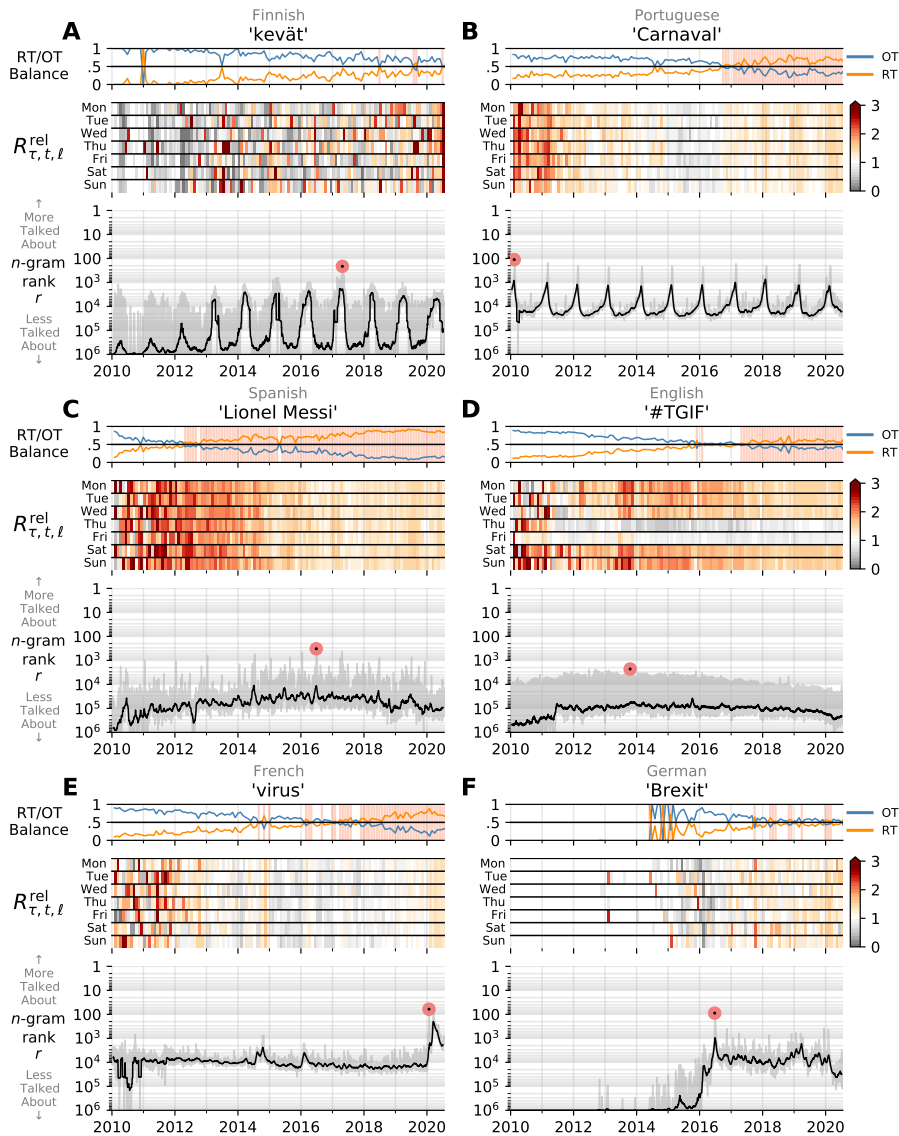We provide Python code for generating arbitrary contagiograms along with further examples at https:

FIG. 3. **Contagiograms: Augmented time series charting the social amplification of n-grams.** In each contagiogram, above the basic n-gram rank time series, the top panel displays the monthly relative usage of each n-gram, $R_{\tau,t,\ell}$ (Eq. (4)), indicating whether they appear organically in new tweets (OT, blue), or in retweeted content (RT, orange). The shaded areas denote months when the balance favors spreading, suggestive of story contagion. The middle (second) panel then shows retweet usage of an n-gram relative to the background rate of retweeting, $R_{\tau,t,\ell}^{\text{rel}}$ (Eq. (5)). **A–B.** The seasonal cycle of the 1-gram 'spring' in Finnish is different than the annual cycle of the word 'Carnaval' in Portuguese. Spring is often mentioned in organic tweets while the balance of the word 'Carnaval' favors retweets exceeding the social contagion threshold starting from 2017. **C.** The time series for 'Lionel Messi' in Spanish tweets exhibits a similar pattern of social amplification as a famous soccer player who is talked about regularly. **D.** The hashtag #TGIF ('Thank God It's Friday') shows a strong weekly cycle, relatively unamplified on Thursday and Friday. **E.** The time series of the 1-gram 'virus' in French shows strong relative retweeting following global news about the early spread of COVID-19 in 2020-01. **F.** We observe mild spikes at the beginning of the German dialog around the withdrawal of the UK from the EU shifting to a even balance of the 1-gram 'Brexit' across organic and retweeted content.

//gitlab.com/compstorylab/contagiograms. The figure-making scripts interact directly with the Storywrangler database, and offer a range of configurations.
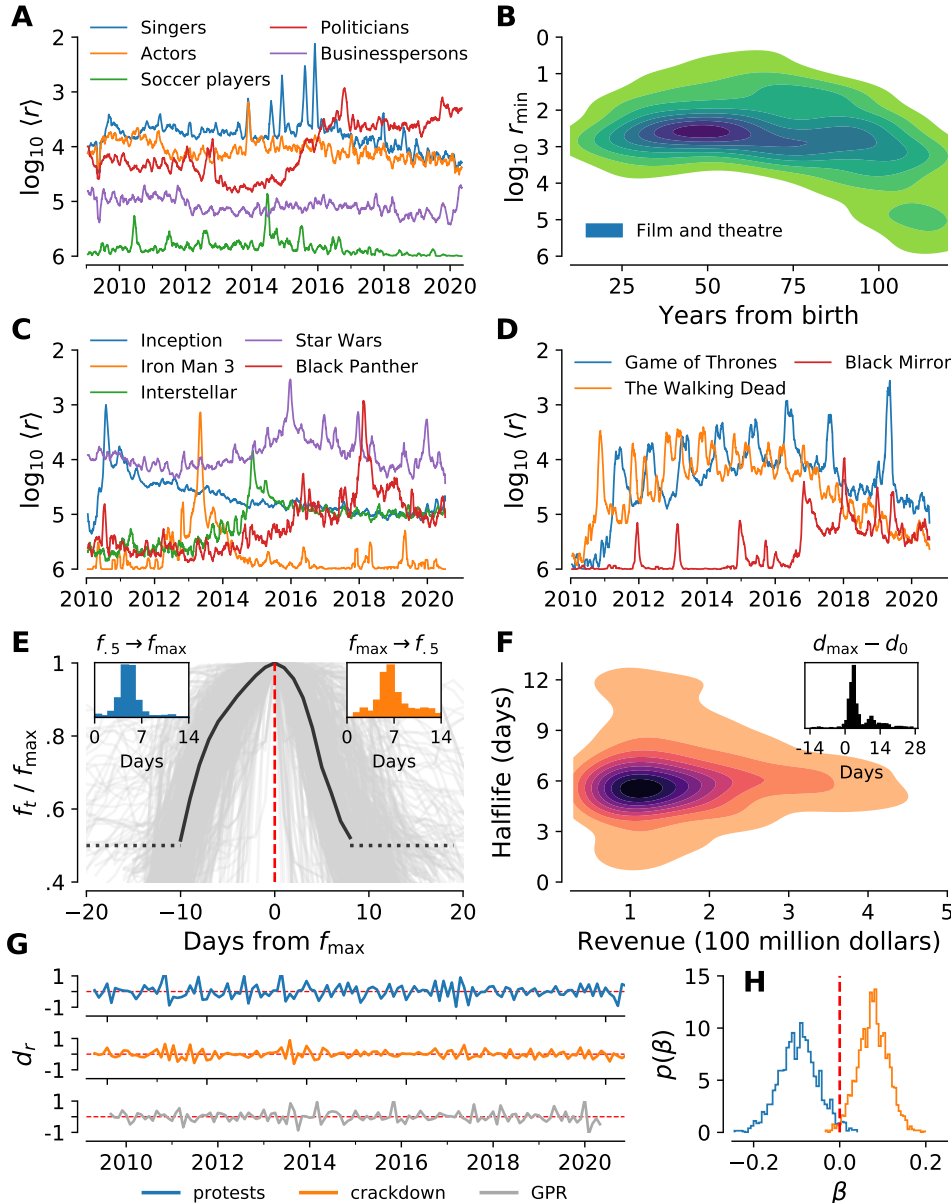
C. Case studies

As a demonstration of our dataset's value across diverse fields, we briefly present three case studies. We analyze (1) The dynamic behavior of famous individuals' full names and their association with the individuals' ages; (2) The relationship between movie revenue and anticipatory dynamics in title popularity; and (3) The potential of social unrest related words to predict future geopolitical risk.

We examine the dialog around celebrities by cross-referencing our English 2-grams corpus with names of famous personalities from the Pantheon data set [40].

We searched through our English n-grams data set and selected names that were found in the top million ranked 2-grams for at least a day between 2010/01/01 and 2020/06/01. In Fig. 4A, we display a monthly rolling average (centered) of the average rank for the top 5 individuals for each category $\langle r_{\min(5)} \rangle$ (see also Fig. S5). In Fig. 4B, we display a kernel density estimation of the top rank achieved by any of these individuals in each industry as a function of the number of years since the recorded year of birth. We note high density of individuals marking their best rankings between 40 and 60 years of age in the film and theatre industry. Different dynamics can be observed in Fig. S5 for other industries.

We next investigate the conversation surrounding major film releases by tracking n-grams that appear in titles for 636 movies with gross revenue above the 95th percentile during the period ranging from 2010/01 to 2017/07 [41]. We find a median value of 3 days post-

FIG. 4. **Three case studies joining Storywrangler with other data sources. A–B.** We cross-reference our English 2-grams corpus with famous figures from the Pantheon data set [40]. **A.** Monthly rolling average of rank $\langle r \rangle$ for the top-5 ranked Americans born in the last century in each category for a total of 960 individuals found in the Pantheon data set. **B.** Kernel density estimation for the highest rank r_{min} achieved by 751 personalities in the film and theater industry as a function of their age. **C–F.** Exploration of movies being talked about on Twitter and box office success. **C.** Rank time series for example movie titles showing anticipation and decay. **D.** Contrasting with **C**, rank time series for TV series titles. In panels **E** and **F**, Time series and half-life revenue comparison for 636 movie titles with gross revenue at or above the 95th percentile released between 2010/01 and 2017/07 [41]. **G–H.** The Storywrangler data set can also be used to potentially predict political and financial turmoil. Percent change in the words 'protests' and 'crackdown' in month m are significantly associated with percent change in a geopolitical risk index in month $m+1$ [42]. We display the percent change time series in panel **G** and distributions of coefficients of a fit linear model in panel **H**. See Secs. S1 E, S1 F, and S1 G for details of each study.

release for peak normalized frequency of usage for movie n-grams (Fig. 4F inset). Growth of n-gram usage from 50% ($f_{.5}$) to maximum normalized frequency (f_{max}) has a median value of 5 days across our titles. The median value of time to return to $f_{.5}$ from f_{max} is 6 days. Looking at Fig. 4E we see the median shape of the spike around movie release dates tends to entail a gradual increase to peak usage, and a relatively more sudden decrease when returning to $f_{.5}$. There is also slightly more spread in the time to return to $f_{.5}$ than compared with the time to increase from $f_{.5}$ to f_{max} (Fig. 4E insets).

In Figs. 4G and H, we show that changes in word popularity can predict future changes in geopolitical risk, which we define here as "a decline in real activity, lower stock returns, and movements in capital flows away from emerging economies" following the US Federal Reserve [42]. We chose a set of words that we *a priori* believed might be associated with geopolitical risk as design variables and a geopolitical index created by the US Federal Reserve as the response. We fit a linear model using the values of the predictors at month m to predict the value of the geopolitical risk index at month $m+1$. Two of the words, 'crackdown' and 'protest', were significantly associated with changes in the geopolitical risk index.

For details about our methodology and further results, see Secs. S1 E, S1 F, and S1 G.

IV. CONCLUDING REMARKS

There are many aspects of Storywrangler to either improve or introduce. For high volume languages, we would aim for higher temporal resolution—at the scale of minutes for, say, English and Spanish—and as would be limited by requiring n-gram counts to exceed some practical minimum. We would also want to expand our language parsing to cover continuous-script languages such as Japanese and Chinese.

Another large space of natural improvements would be to categorize tweets in ways other than by language identification such as geography, user type (e.g., people, institutions, or bots), and topic (e.g., all tweets containing 'Trump' or 'Brexit'). We note that for Twitter, features like location and user type are more difficult to establish with as much confidence as we have in language identification.

Topic-based subsets are particularly promising as they would allow for explorations of language use, ambient framings, narratives, and conspiracy theories. In this initial version of Storywrangler, 2-grams and 3-grams would make possible certain analyses of the temporal evolution of 1-grams adjacent to an anchor 1-gram or 2-gram. By using wild cards, linguists will in principle be able to track patterns of popular language use in a way that Google Books n-gram corpus seems but fails to do [1, 2]. Similarly, journalists and political scientists could chart 1-grams being used around, for example, 'Trump' or '#BlackLivesMatter' over time.

Looking outside of text, a major possible expansion of the data set would be to incorporate images and video, which have evidently both become increasingly integral to social media over the last decade. And moving away from Twitter, we could develop similar n-gram data sets for other platforms where social amplification is a recorded feature (e.g., Reddit).

We add a caution against cherry picking time series. The example time series we show in Fig. 2 and through our site storywrangling.org do evidently track historical events through the complex sociotechnical reporting system that is Twitter. Still, there is potential for misuse of the data, whether purposefully or not, to portray via an isolated time series evidence of a particular story or sociocultural evolution. Words and phrases drift in meaning and other terms take their place. For example, 'coronavirus' gave way to 'covid' as the dominant term of reference on Twitter for the COVID-19 pandemic in the first six months of 2020. To in part properly demonstrate a trend, researchers would need to at least marshal together thematically related n-grams, and do so in a data-driven way, as we have attempted to do for our case studies in Sec. III C. Thoughtful consideration of overall and normalized frequency of usage would also be needed to show whether a topic is changing in real volume.

In building Storywrangler, our primary goal has been to curate and share a rich, language-based ecology of interconnected n-gram time series derived from Twitter. Unlike most text corpora, our n-gram data set contends with popularity, allowing for the examination of story amplification, and we emphasize the importance of using contagiograms as visualization tools that go beyond presenting simple time series. We see some of the strongest potential for future work in the coupling of Storywrangler with other data streams to enable, for example, data-driven, computational versions of journalism, linguistics, history, economics, and political science.

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman, Science Magazine **331**, 176 (2011).

[2] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, PLoS ONE **10**, e0137041 (2015).

[3] H. Christenson, Library Resources & Technical Services **55**, 93 (2011).

[4] M. Gerlach and F. Font-Clos, Entropy **22**, 126 (2020).

[5] E. Sandhaus, "The New York Times Annotated Corpus," Linguistic Data Consortium, Philadelphia (2008).

[6] D. Beeferman, W. Brannon, and D. Roy, arXiv preprint arXiv:1907.07073 (2019).

[7] L. Hollink, A. Bedjeti, M. van Harmelen, and D. Elliott, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (2016) pp. 1377–1382.

[8] W. Mieder, *Proverbs: A handbook* (Greenwood Publishing Group, 2004).

[9] J. Abello, P. Broadwell, and T. R. Tangherlini, Communications of the ACM **55**, 60 (2012).

[10] T. R. Tangherlini and P. Leonard, Poetics **41**, 725 (2013).

[11] Q.-H. Vuong, Q.-K. Bui, V.-P. La, T.-T. Vuong, V.-H. T. Nguyen, M.-T. Ho, H.-K. T. Nguyen, and M.-T. Ho, Palgrave Communications **4**, 1 (2018).

[12] J. T. Woolley and G. Peters, Santa Barbara, CA. (2008), accessed on 2019/09/15.

[13] R. B. Primack, H. Higuchi, and A. J. Miller-Rushing, Biological Conservation **142**, 1943 (2009).

[14] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, Science Advances **6**, eaay3539 (2020).

[15] C. S. S. Peirce, The Monist , 492 (1906).

[16] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, patterns (Addison-Wesley, Cambridge, MA, 1949).

[17] J. Merritt, *Learning to speak God from scratch: Why sacred words are vanishing—and how we can revive them* (Convergent Books, 2018).

[18] J. Bohannon, "Google opens books to new cultural studies," (2010).

[19] A. Koplenig, Digital Scholarship in the Humanities **32**, 169 (2017).

[20] A. Koplenig, Digital Scholarship in the Humanities **32**, 159 (2017).

[21] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, Journal of Computational Science **21**, 24 (2017), available online at http://arxiv.org/abs/1503.03512.

[22] S. Hong and D. Nadler, in *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (2011) pp. 182–186.

[23] A. Younus, M. A. Qureshi, F. F. Asar, M. Azam, M. Saeed, and N. Touheed, in *2011 International Conference on Advances in Social Networks Analysis and Mining* (IEEE, 2011) pp. 618–623.

[24] E. M. Cody, A. J. Reagan, P. S. Dodds, and C. M. Danforth, ArXiv **abs/1608.02024** (2016).

[25] T. Sakaki, M. Okazaki, and Y. Matsuo, in *Proceedings of the 19th international conference on World wide web* (2010) pp. 851–860.

[26] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, Science **334**, 509 (2011).

[27] H. Gao, G. Barbier, and R. Goolsby, IEEE Intelligent Systems **26**, 10 (2011).

[28] V. Lampos and N. Cristianini (2010) pp. 411 – 416.

[29] A. Culotta, in *Proceedings of the First Workshop on Social Media Analytics*, SOMA 10 (Association for Computing Machinery, New York, NY, USA, 2010) p. 115122.

[30] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler, EPJ Data Science **4**, 19 (2015).

[31] J. Mellon and C. Prosser, Research & Politics **4**, 2053168017720008 (2017).

[32] W. Liu and D. Ruths, in *2013 AAAI Spring Symposium Series* (2013).

[33] R. Cohen and D. Ruths, in *Seventh international AAAI conference on weblogs and social media* (2013).

[34] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras, PloS one **10**, e0138717 (2015).

[35] X. Zheng, J. Han, and A. Sun, IEEE Transactions on Knowledge and Data Engineering **30**, 1652 (2018).

[36] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds, "The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020," (2020), available online at http://arxiv.org/abs/2003.03667.

[37] "FastText language identification," https://fasttext.cc/docs/en/language-identification.html (2017).

[38] P. S. Dodds *et al.*, "Long-term word frequency dynamics derived from twitter are corrupted: A bespoke approach to detecting and removing pathologies in ensembles of time series," (2020).

[39] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth, "Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop," (2019), available online at https://arxiv.org/abs/1910.00149.

[40] A. Z. Yu, S. Ronen, K. Hu, T. Lu, and C. A. Hidalgo, Scientific data **3**, 150075 (2016).

[41] F. M. Harper and J. A. Konstan, Acm transactions on interactive intelligent systems (tiis) **5**, 1 (2015).

[42] D. Caldara and M. Iacoviello, FRB International Finance Discussion Paper (2018).

[43] Q. Ke, Y.-Y. Ahn, and C. R. Sugimoto, PLOS ONE **12**, 1 (2017).

[44] H. A. Simon, Biometrika **42**, 425 (1955).

[45] D. D. S. Price, Journal of the American Society for Information Science , 292 (1976).

[46] Https://gitlab.com/compstorylab/storywrangler.

[47] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, Nature Scientific Reports **5**, 12209 (2015).

[48] E. Chenoweth and M. J. Stephan, Foreign Aff. **93**, 94 (2014).

[49] M. D. Hoffman and A. Gelman, Journal of Machine Learning Research **15**, 1593 (2014).

[50] A. Gelman, D. B. Rubin, *et al.*, Statistical science **7**, 457 (1992).

[51] Twitter, "Developer application program interface (API)," https://developer.twitter.com/en/docs/ads/campaign-management/api-reference (2019), [Online; accessed 01-October-2019].

S1. SUPPLEMENTARY MATERIAL: METHODS

A. Language classification

Document language identification and detection (LID) is a well developed area of natural language processing (NLP). However, most tools have limitations when analyzing short text such as tweets. Here, we use the state-of-the-art algorithm FastText [37] to classify tweets, fully articulating and examining this choice in Ref. [36].

Before classifying tweets, we filter out Twitter-specific content such as hashtags, handles, and HTML codes, along with links, emojis, and any redundant whitespaces. This simple procedure allows us to filter out potential implicit classification biases as most of them tend to be shared in English (e.g., emojis).

B. Social amplification and contagion

Twitter enables social amplification on the platform through the use of retweets and, from 2015 on, quote tweets. Users—including the general public, celebrities, scientists, decision-makers, and social bots [43]—can intervene in the information spread process and amplify the volume of any content being shared. We categorize tweets into two major classes: organic tweets (OTs) and retweets (RTs) as described in Ref. [36]. Organic tweets represent the set of new information being shared on the platform on a daily bases, whereas the RTs reflect the volume of information being socially amplified on Twitter. During that process, we enrich the text body of RTs with (RT @userHandle: ...) to indicate the original user of the retweeted text. Our categorization enables users of the Storywrangler data set to tune the amplification processes of the rich-get-richer mechanism [44, 45] by dialing the ratio of RTs added to the n-grams corpus.

C. Twitter n-grams

For our initial version of Storywrangler for Twitter, we have extracted n-grams from tweets where $n \in \{1, 2, 3\}$. We record raw n-gram frequency (or count) at the day scale for each language (including unidentified), and for Twitter as a whole.

A 1-gram is a continuous string of characters bounded by either whitespace or punctuation marks. For example, the word 'the' is one of the prominent 1-grams in English. The 2-gram 'here?' consists of the 1-grams: 'here' and '?'. Numbers and emojis also count as 1-grams.

Similarly, a word bound by two quotes (e.g., "sentient") would be a 3-gram, and the expression 'see the light' is a 3-gram, and so forth.

To accommodate the rich lexicon of Twitter, we have designed a special n-gram parser. We parse currency (e.g., $9.99), floating numbers (e.g., 3.14), and date/time strings (e.g., 2001/9/11, 2018-01-01, 11:59pm) all as 1-grams. We curate links (e.g., https://www.google.com/), handles (e.g., @NASA), and hashtags (e.g., #metoo) as 1-grams. We endeavor to combine contractions and acronyms as single objects and parse them out as 1-grams (e.g., "It's", "well-organized", and "B&M").

Emojis are uniquely and interestingly complex entities. People-centric emojis can be composed of components such as skin-tone modifiers, hair-type modifiers, and family structures. Emoji flags are all two component objects. The most elaborate emojis are encoded by seven or more unicode elements, rendering them difficult to extract as single entities. After contending with many emoji-parsing problems, we record all emojis as 1-grams. We consider repeated emojis with no intervening whitespace—a common feature in tweets—to be a series of 1-grams.

In Fig. S1, we show contagiograms for 12 example n-grams that involve punctuation, numbers, handles, hashtags, and emojis. A few more examples across various languages can be seen in Fig. S2.

Our n-gram parser is case sensitive. For example, search queries made on storywrangling.org for 'New York City' and a search for 'new york city' would return different results. Normalized frequencies and rankings in our daily Zipf distributions are consequently for case-sensitive n-grams.

Although we can identify tweets written in continuous-script-based languages (e.g., Japanese, Chinese, and Thai), our current parser does not support breaking them into n-grams. We label tweets as Undefined (und) to indicate tweets that we could not classify with a confidence score above 25%. The resulting n-grams are allocated to an "Undefined" category as well as to the overall Twitter n-gram data set.

To enable access to our dataset, we maintain a MongoDB database of the one million most frequently used n-grams on each day for each language. We index these collections by date, to allow efficient queries for all n-grams on a given day, as well as by n-gram, which allows for rapid time series queries. Data is typically inserted within two days, i.e., counts from Monday will be available by midnight Wednesday.

Our source code along with our documentation is publicly available online on a Gitlab repository [46].

D. Constructing daily Zipf distributions

For ease of usability, we maintain two sets of daily measurements for each n-gram in our data set: raw frequency (count), normalized frequency (probability), and tied rank with and without retweets included. We make the default ordering for the Zipf distribution files according to usage levels of n-grams for all of a given language on Twitter (i.e., including all retweets and quote tweets). Again, all daily distributions are made according to UTC calendar days.
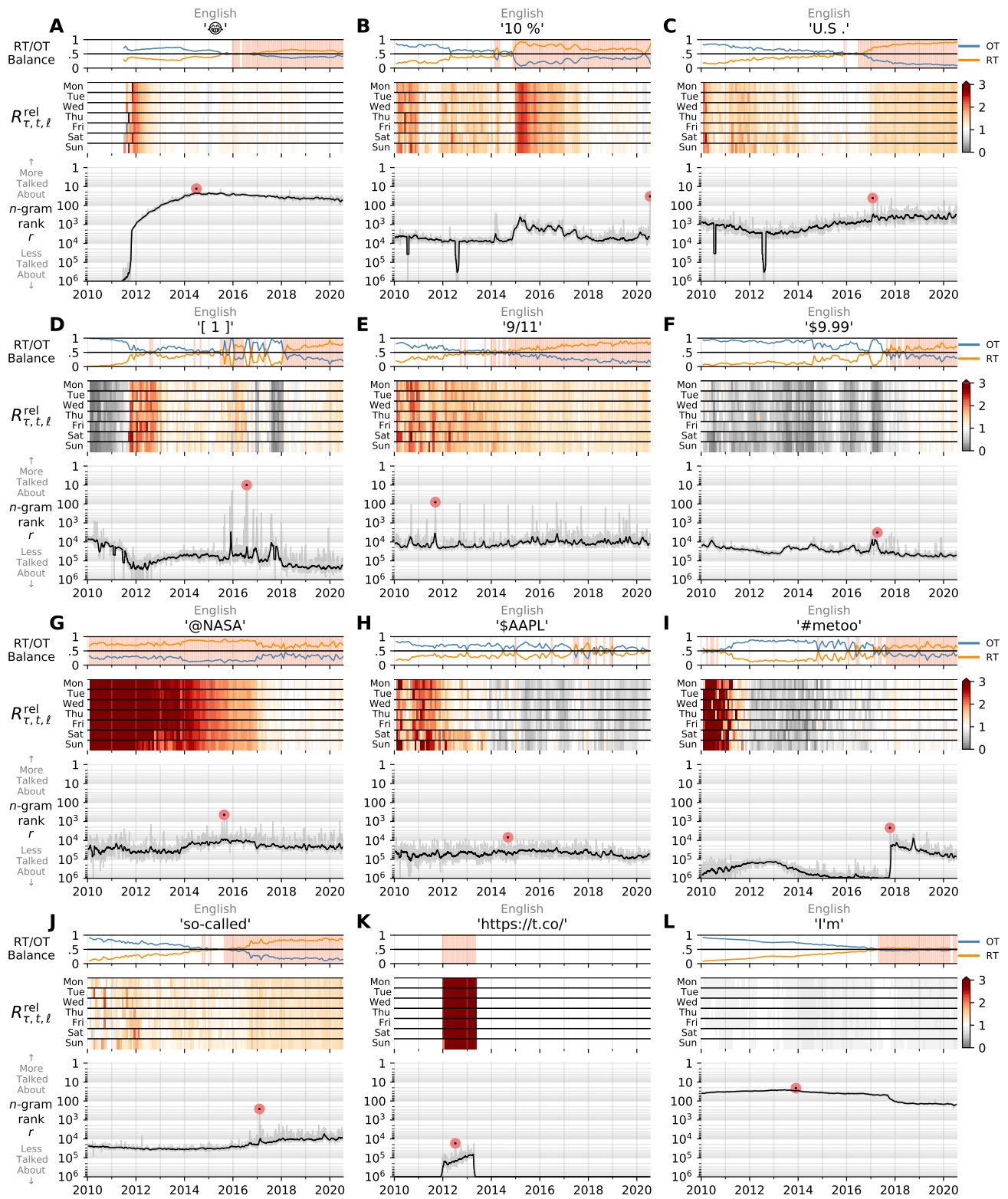
FIG. S1. Example contagiograms for Twitter n-grams involving emojis, punctuation, numerals, and so on.
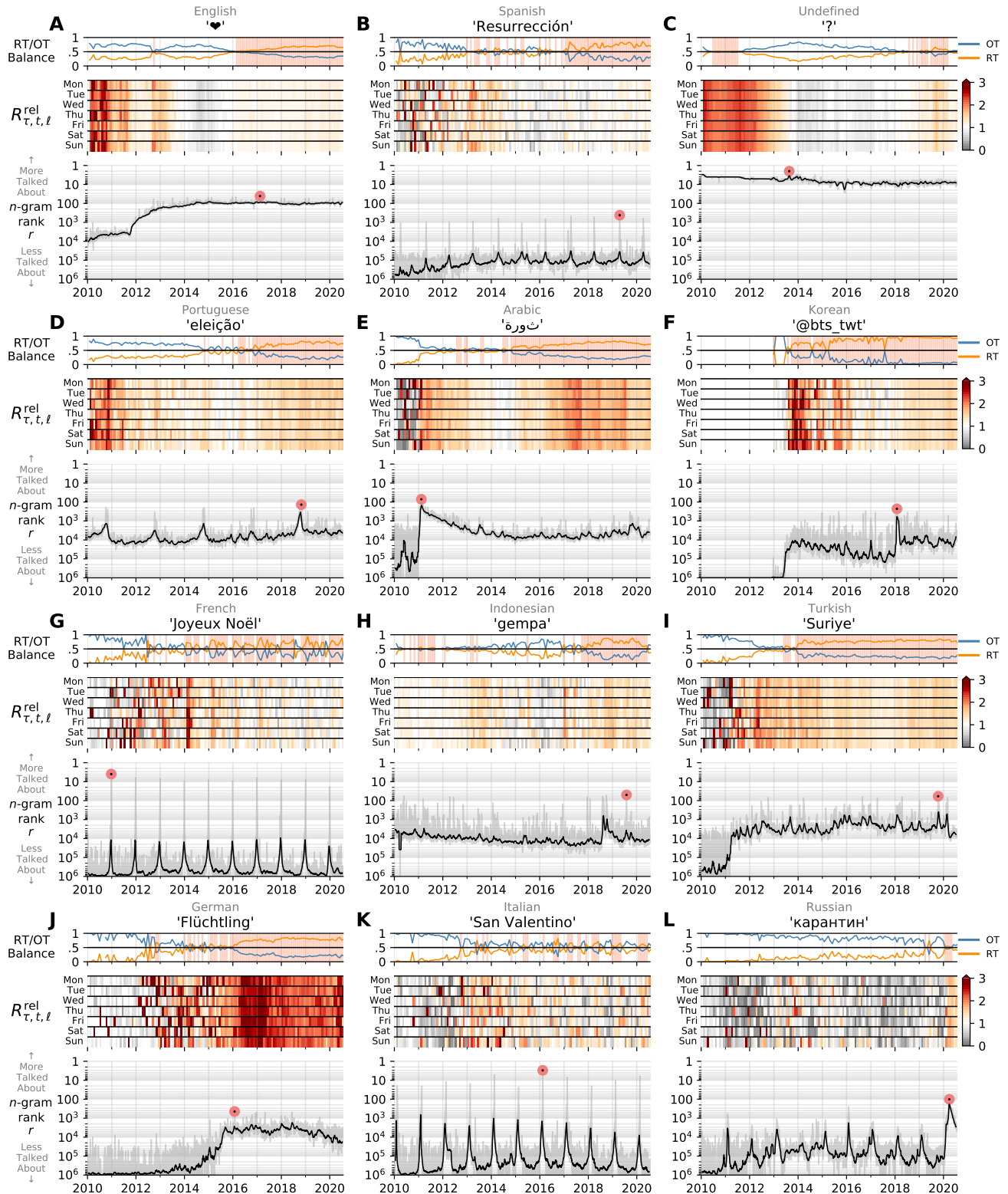
FIG. S2. **The interplay of social amplification across various languages.** We observe a wide range of sociotechnical dynamics starting with n-grams that are often mentioned within OTs and RTs equivalently to others that spread out after a geopolitical event and more extreme regimes whereby some n-grams are consistently amplified. English translations of n-grams: **B.** 'Resurrection', **D.** 'election', **E.** 'revolution', **G.** 'Merry Christmas', **H.** 'earthquake', **I.** 'Syria', **J.** 'Refugee', **K.** 'Saint Valentine', and **L.** 'quarantine'.
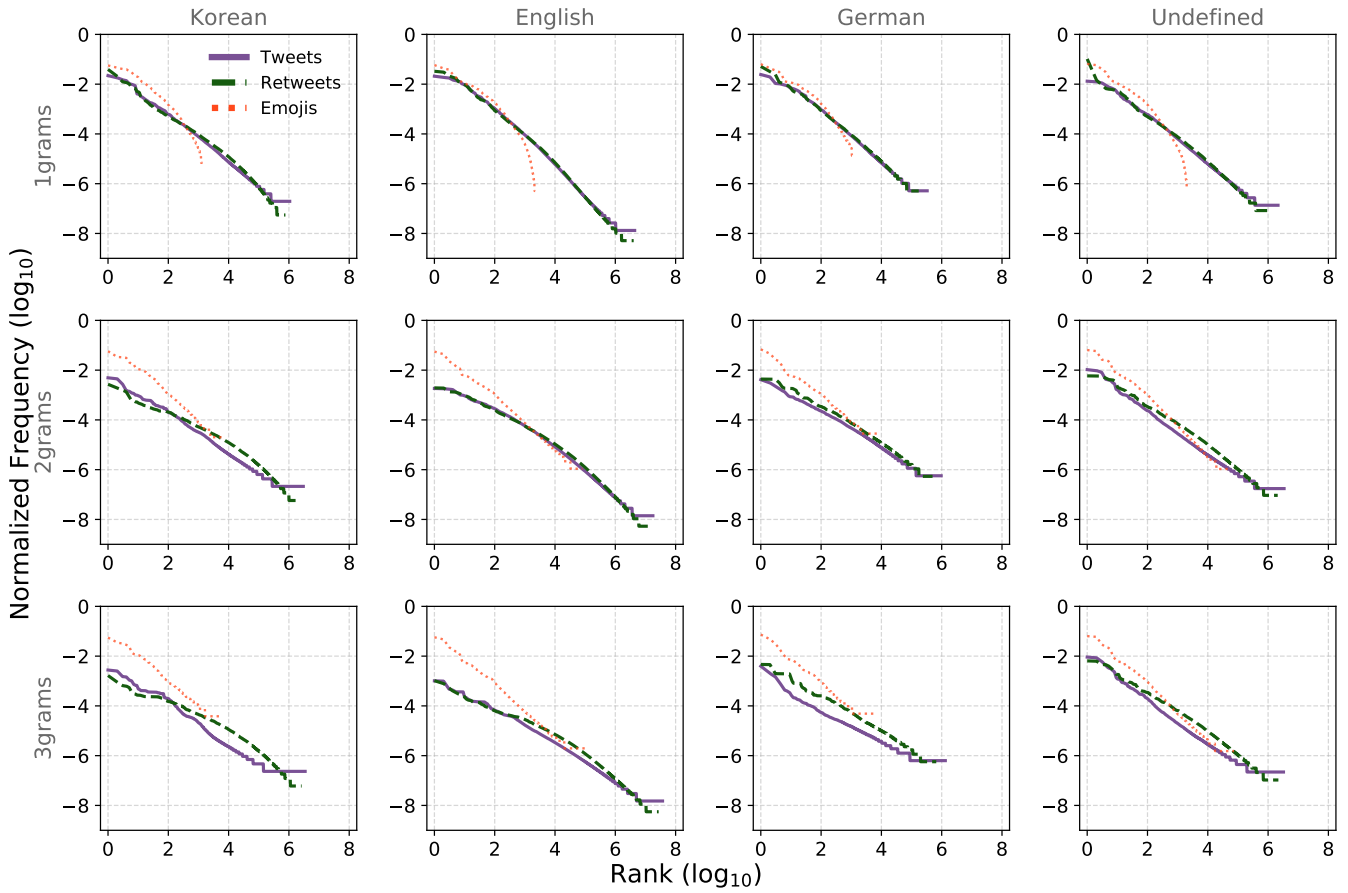
FIG. S3. **Zipf distributions for Korean, English, German and undefined language categories for October 16, 2019 on Twitter.** "Tweets" refer to organic content, "retweets" retweeted content, and "emojis" are ngrams comprised of strictly emojis (organic and retweets combined).

We compute relative daily rate of usage by dividing total number of occurrences of a given n-gram by the total number of n-grams for that day. We then rank all n-grams for a given day to create daily Zipf distribution [16] for all languages in our data set. If two or more distinct n-grams have the same number of instances (raw frequency), then we adjust their ranks by taking the average rank (i.e., tied-rank).

We do not mix n-grams for different values of n, and leave this as an important future upgrade [47]. Users of the viewer storywrangling.org, will need to keep this in mind when considering time series of n-grams for different n. In comparing, for example, 'NYC' (a 1-gram) to 'New York City' (a 3-gram), the shapes of the curves will be meaningful while the ranks (or raw frequencies) of the 1-gram and 3-gram will not be.

We show complementary cumulative distribution functions (CCDFs) of OTs, RTs and emojis for 1-, 2-, and 3-grams in Fig. S3 and Fig. S4.

E. Pantheon case study

We examine the dialog around celebrities by cross-referencing our English 2-grams corpus with names of famous personalities from the Pantheon data set [40]. The data set has over 10 thousand biographies. We use the place and date of birth to select Americans born in the last century.

We searched through our English n-grams data set and selected names that were found in the top million ranked 2grams for at least a day between 2010/01/01 and 2020/06/01. Our list contains 1010 individuals. We show the average best rank $\bar{r}_{\min}$, median best rank $\tilde{r}_{\min}$, and best rank $r^*_{\min}$ for all individuals in each occupation in Tab. S1. In Figs. S5A and B, we display a monthly rolling average (centered) of the average rank for the top 5 individuals for each category $\langle r_{\min(5)} \rangle$.

Additionally, we select a total of 1162 celebrities that were also found in the top million ranked 2grams for at least a day between 2010/01/01 and 2020/06/01 in the a few industries (see Tab. S2)

For each of these individuals, we track their age and top daily rank of their names (first and last name). In
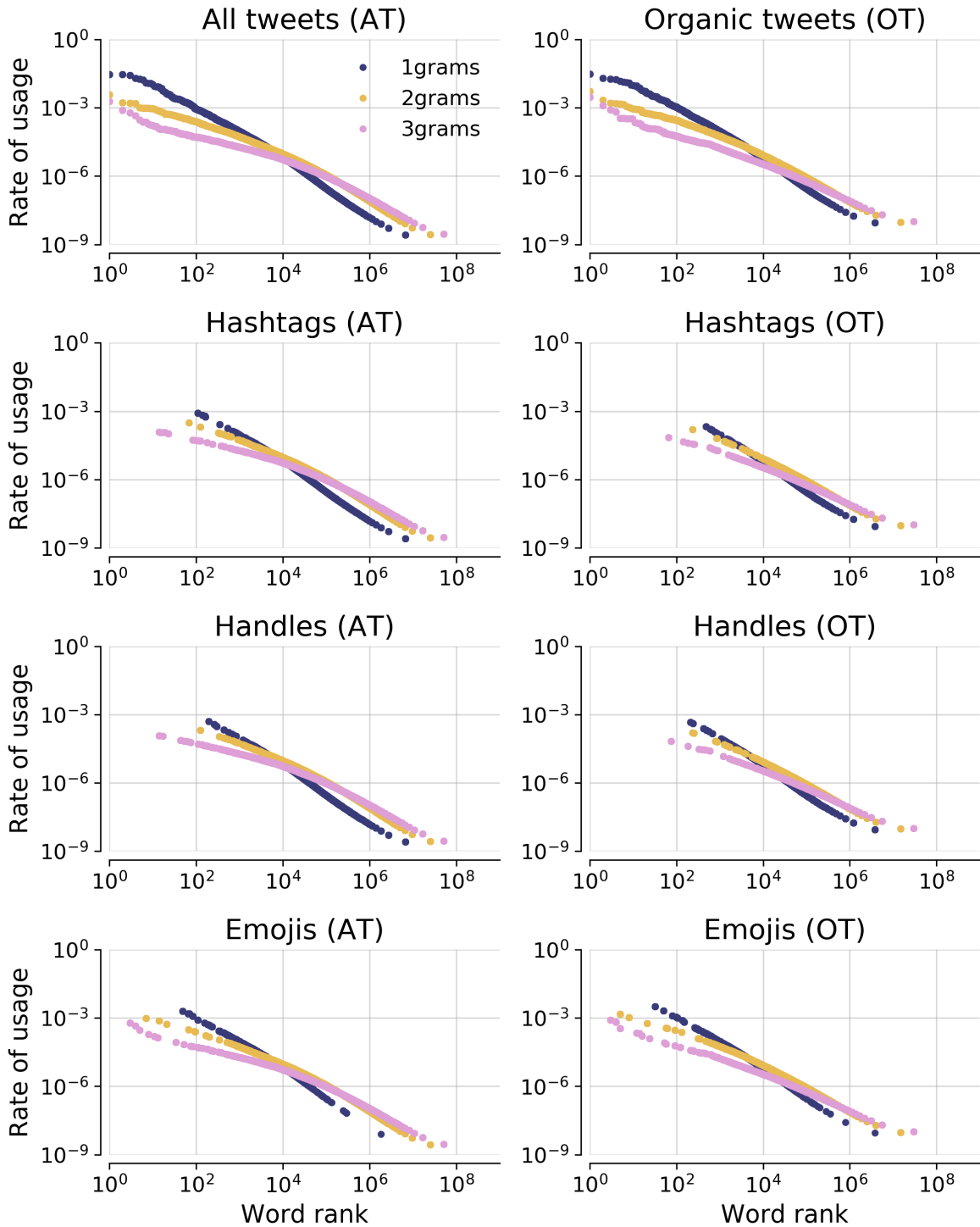
FIG. S4. **Zipf distributions for English on May 1st, 2020.** We show a weighted 1% random sample of 1-grams (blue), 2-grams (yellow), and 3-grams (pink) in all tweets (AT) and organic tweets (OT) accordingly. On the vertical axis, we plot the relative rate of usage of each n-gram in our random sample whereas the horizontal axis displays the rank of that n-gram in the English corpus of that day. We first display Zipf distributions for all n-grams observed in our sample in the first row. We also demonstrate the equivalent distributions for hashtags (second row), handles (third row), and emojis (last row).
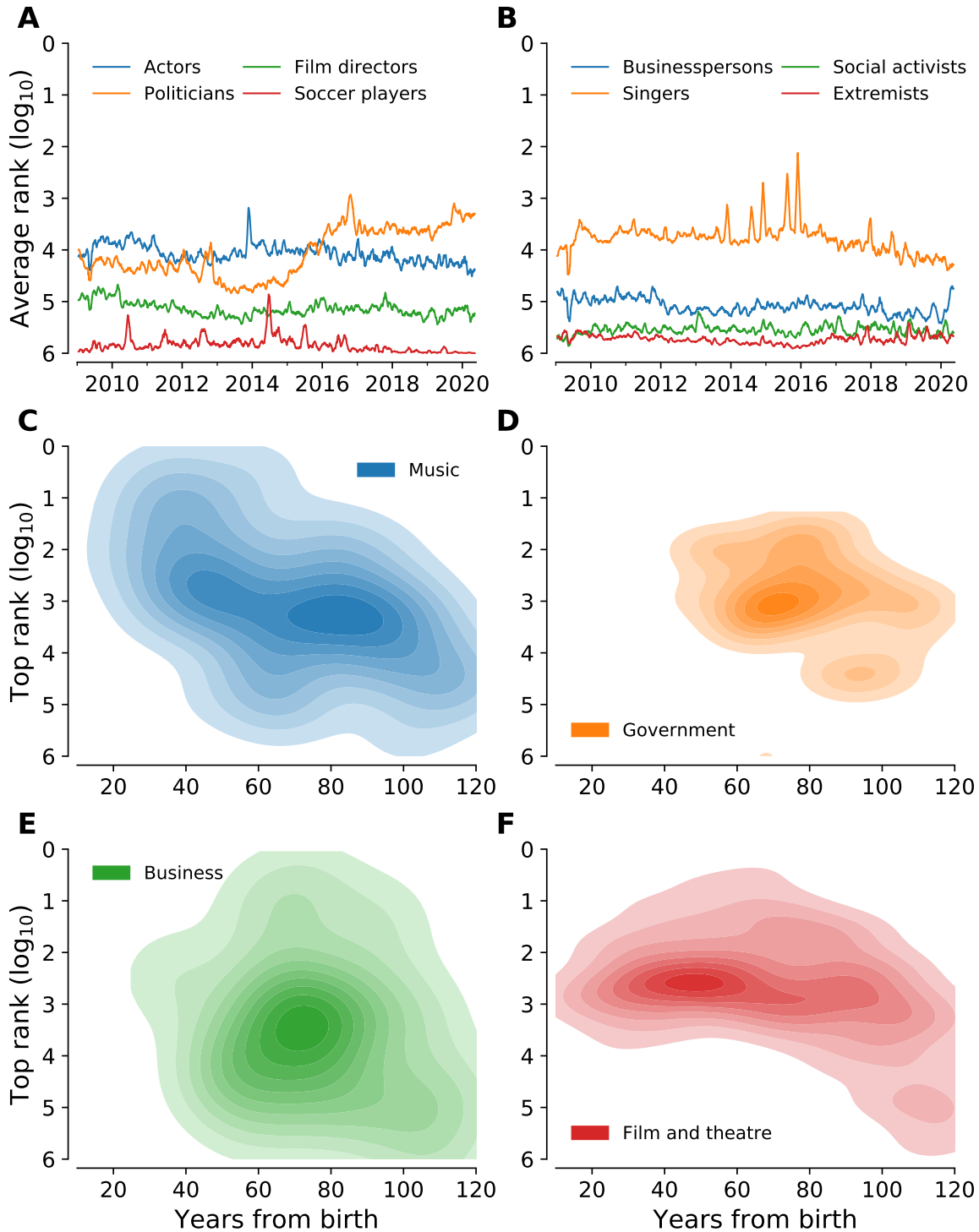
FIG. S5. **Rankings of celebrities on Twitter**. We take a closer look at rankings of famous figures by cross-referencing our English corpus with names of celebrities from the Pantheon dataset [40]. We use their first and last name to search through our 2-grams data set. We select names of Americans who were born in the last century and can be found in the top million ranked 2grams for at least a day between 2010/01/01 and 2020/06/01. In panels **A** and **B**, we display a centered monthly rolling average of the average rank for the top 5 individuals for each category $\langle r_{\min(5)} \rangle$. We also plot the kernel density estimation of the best rank achieved by another 1162 famous characters in each of the following industries: **C.** music, **D.** government, **E.** business, and **F.** film.

Occupation	n	$\bar{r}_{\min}$	$\tilde{r}_{\min}$	$r^*_{\min}$
Actors	674	40,632	9,255	2
Singers	162	59,713	3,479	4
Politicians	59	6,365	1,376	6
Film directors	57	75,783	10,580	13
Business-persons	26	35,737	4195	15
Soccer players	12	20,868	8,507	25
Social activists	10	20,302	1,781	841
Extremists	10	104,621	20,129	117

TABLE S1. Celebrities by occupation

Industry	Individuals
Film and theater	751
Music	324
Government	59
Business	28

TABLE S2. Celebrities by industry

Figs. S5A, B, C, and D, we display kernel density estimation of the top rank achieved by any of these individuals in each industry as a function of the number of years since the recorded year of birth (age of the cohort).

F. Movies case study

We investigate the conversation surrounding major film releases by tracking n-grams that appear in movies titles. From the MovieLens data set [41], we selected 636 movies with gross revenue above the 95th percentile during the period ranging from 2010/01 to 2017/07. We then retrieved the normalized frequency time series for up to the first 3-grams of a movie's title (e.g., "Prince of Persia: The Sands of Time" would correspond to the 3gram time series "Prince of Persia"). From there, we look for the maximum daily normalized frequency. To disambiguate between movies within the same franchise and/or titles with common n-grams, we restrain this search to the release year of the given movie. With the peak usage in the year of a movie's release, we then search backward for the date on which the n-gram usage first breaks 50% of the peak usage normalized frequency, $f_{.5}$. Similarly we search forward, from peak usage, for the date on which the time series first declines below $f_{.5}$.

Peak conversation surrounding major movies tends to occur a few days after the release data of the title. We find a median value of 3 days post-release for peak normalized frequency of usage for movie n-grams (Fig. 4F inset). Growth of n-gram usage from 50% to maximum normalized frequency has a median value of 5 days across our 636 titles. The median value of time to return to 50% from maximum normalized frequency is 6 days. Looking at Fig. 4E we see the median shape of the spike around movie release dates tend to entail a gradual increase to peak usage, and a more sudden decrease when returning

to 50% of maximum normalized frequency. There is also slightly more spread in the time to return to 50% normalized frequency of usage than compared with the time to increase from 50% to maximum usage (Fig. 4E insets).

G. Geopolitical risk case study

Twitter sentiment has already been shown to provide a useful signal in monitoring public opinion [22–24]. The aggregation process in which individual tweet documents are turned into popularity time series reduces the time and computation power required to use this data in models of political sentiment and public opinion. As a case study, we sought to predict values of a geopolitical risk index (GPR) using popularity of words that we heuristically associated with (inter)national unrest and popular discontent.

The geopolitical risk index is developed by the U.S. Federal Reserve (central bank) [42]. We chose the words "revolution", "rebellion", "uprising", "coup", "overthrow', "unrest", "protests", and "crackdown" to include as predictors. Since we could not reject the null hypothesis that at least one of the logarithm of normalized frequency time series associated with these words contained a unit root (ADF("overthrow") $= -1.61$, $p = 0.474$), we computed the difference of the logarithm of normalized frequency time series and used these observations as features. We could not reject the null hypothesis that the geopolitical risk time series contained a unit root (ADF $= -0.65$, $p = 0.858$) and therefore sought to predict the log difference of the GPR. Because GPR is computed at monthly frequency, we resampled normalized word frequencies to monthly normalized frequency by taking the average of the lagged month's values. For example, the normalized frequency of the word "crackdown" sampled at month level timestamped at 2010/03/31 was taken to be the average daily normalized frequency of the word "crackdown" from 2010/03/01 through 2010/03/31.

We fit a linear model that hypothesized a linear relationship between the log difference in normalized word frequencies for each of the words listed in the previous paragraph and the log difference in GPR. The likelihood function of this model took the form

$$p(\log_e \Delta \vec{y} | \vec{\beta}, \sigma) = \prod_{t=0}^{T-1} \text{Normal}(\log_e \Delta \vec{X}_t \vec{\beta}, \sigma^2). \quad (6)$$

We denote $\vec{y} \equiv (y_1, ..., y_T)$ and $\Delta y_t \equiv y_t - y_{t-1}$, while $\vec{X}$ is a $N \times (p+1)$ matrix that is the design matrix of the linear model. It contains the log difference in normalized word frequencies and a preprended vector with all values equal to one. We performed regularization by placing a zero-mean normal prior on the vector of coefficients as $\vec{\beta} \sim$ MultivariateNormal($\vec{0}, \vec{I}$) with $\vec{0}$ the $p + 1$-dimensional zero vector and $\vec{I}$ the $(p+1) \times (p+1)$-dimensional identity matrix. We placed a weakly informative prior on the
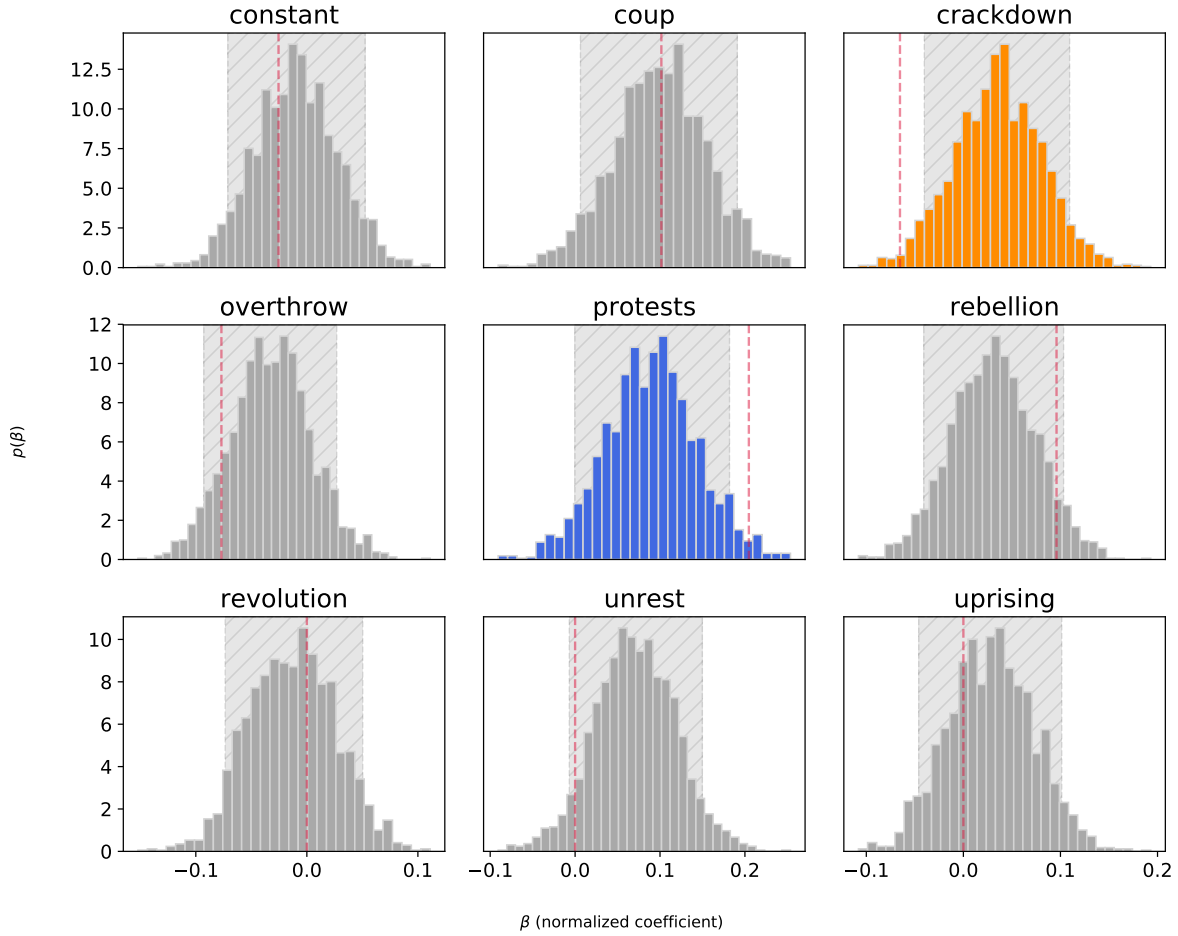
FIG. S6. Percent change in word popularity is significantly associated with percent change in a Bayesian linear model future geopolitical risk (GPR) index level for two words, "crackdown" and "protests", out of a panel of eight words: "coup", "crackdown", "overthrow", "protests", "rebellion", "revolution", "unrest", and "uprising". The sign of the coefficient differs between the words, with "crackdown" positively associated with GPR and "protests" negatively associated. We hypothesize that this is because a crackdown is a violent response to a social movement by definition, while protests do not necessarily entail violence and can serve as a peaceful outlet for frustration [48].

noise scale, $\sigma \sim \mathrm{LogNormal}(0,1)$, as we were *a priori* unsure of what noise level the data would exhibit.

We sampled from the model using the No U-Turn Sampler (NUTS) algorithm [49] for 500 warmup iterations and 2000 iterations of sampling. There is strong evidence to suggest that the sampler converged since the maximum Gelman-Rubin statistic [50] for all priors was less than 1.01 (max $\hat{R} = 1.0009$).

We computed centered 90% credible intervals for each of the model coefficients β_k, $k = 0, ..., p$. (A centered $Q\%$ credible interval for the univariate random variable $Y \sim p(y)$ is an interval (a, b) such that $\frac{1}{2}(1 - \frac{Q}{100}) = \int_{-\infty}^{a} dy \ p(y) = \int_{b}^{\infty} dy \ p(y)$. For example, a centered 90% credible interval is an interval such that $0.05 = \int_{-\infty}^{a} dy \ p(y) = \int_{b}^{\infty} dy \ p(y)$.) We termed a relationship significant if the 90% credible interval did not contain zero. Log difference in normalized usage frequency of

"crackdown" was significantly positively associated with future log difference in GPR, while log difference in normalized usage frequency of "protests" was significantly negatively associated with future log difference in GPR. We speculate that this is because increases in the usage of "crackdown" mean that a popular revolt is already in the process of being crushed, a realization of increased geopolitical uncertainty, while increases in the usage of "protests" means that citizens are able to peacefully vent grievances without resorting to violence, lessening the possibility of future unrest.

S2. SUPPLEMENTARY MATERIAL: DATA SET

A. Twitter data set basics

From 2008/09/09 on, we have been collecting a random subset of approximately 10% of all public messages using Twitter's Decahose API [51]. Every day, half a billion messages are shared on Twitter in dozens of languages. As of 2020/06, our data collection comprised around 150 billion messages, requiring 100TB of storage. We group tweets by day according to Coodinated Universal Time (UTC).

	AT	**OT**	**RT**	Unique
μ	3.71×10^8	1.73×10^8	1.98×10^8	1.68×10^7
σ	1.97×10^8	8.89×10^7	1.45×10^8	7.83×10^6
min	1.71×10^6	1.39×10^6	1.54×10^3	1.90×10^5
25$^{\text{th}}$	2.05×10^8	1.39×10^8	5.49×10^7	1.17×10^7
50$^{\text{th}}$	4.35×10^8	1.69×10^8	2.06×10^8	2.01×10^7
75$^{\text{th}}$	5.01×10^8	2.46×10^8	2.82×10^8	2.22×10^7
max	8.95×10^8	3.83×10^8	6.26×10^8	2.75×10^7

TABLE S3. Daily summary statistics for 1-grams

	AT	**OT**	**RT**	Unique
μ	3.57×10^8	1.66×10^8	1.91×10^8	6.30×10^7
σ	3.02×10^8	2.03×10^8	1.48×10^8	3.00×10^7
min	1.59×10^6	1.28×10^6	1.47×10^3	8.26×10^5
25$^{\text{th}}$	1.90×10^8	1.28×10^8	5.22×10^7	4.81×10^7
50$^{\text{th}}$	4.05×10^8	1.55×10^8	1.94×10^8	7.26×10^7
75$^{\text{th}}$	4.68×10^8	2.25×10^8	2.69×10^8	7.89×10^7
max	1.18×10^{10}	9.36×10^9	2.40×10^9	4.59×10^8

TABLE S4. Daily summary statistics for 2-grams

	AT	**OT**	**RT**	Unique
μ	3.74×10^8	1.86×10^8	1.88×10^8	1.15×10^8
σ	5.03×10^8	3.82×10^8	1.70×10^8	5.93×10^7
min	1.47×10^6	1.18×10^6	1.40×10^3	1.19×10^6
25$^{\text{th}}$	1.80×10^8	1.19×10^8	5.19×10^7	9.23×10^7
50$^{\text{th}}$	3.81×10^8	1.43×10^8	1.86×10^8	1.26×10^8
75$^{\text{th}}$	4.37×10^8	2.13×10^8	2.49×10^8	1.43×10^8
max	1.17×10^{10}	1.05×10^{10}	3.45×10^9	5.78×10^8

TABLE S5. Daily summary statistics for 3-grams

B. Detailed data set statistics

It is worth noting again that this is an approximate daily leaderboard of language usage and word popularity. Raw frequencies of exceedingly rare words are roughly one-tenth of the true values with regards to all of Twitter, however, rankings are likely to be subject to change. While our API only serves the top million-ranked n-grams, the complete data set is available upon request.

Our Twitter corpus contains an average of roughly 17 million unique 1-grams every day with a maximum of a little over 27 million unique 1-grams captured on 2013/08/07. Because of the combinatorial properties of language, the numbers of unique 2- and 3-grams strongly outweigh the number unique 1-grams. On average, we extract around 63 million unique 2-grams and 115 million unique 3-grams for each day. On 2012/08/03, we recorded a high of $\sim$ 459 million unique 2-grams, and on 2012/01/03, a high of 578 million unique 3-grams.

We emphasize that these maxima for n-grams reflect only our data set, and not the entirety of Twitter. We have no guarantee that our 10% sample is in fact 10%, and cannot make assertions about the size of Twitter's user base or message volume. Indeed, because we do not have knowledge of Twitter's overall volume (and do not seek to per Twitter's Terms of Service), we deliberately focus on ranks and relative usage rates for n-grams away from the tails of their distributions.

In Tab. S3, we show daily summary statistics for 1-grams broken by each category in our data set. We demonstrate the same statistical information for 2-grams and 3-grams in Tab. S4 and Tab. S5 respectively. We show a time series of the unique number of n-grams captured daily in Fig. S7 and the statistical distributions of each category in Fig. S2 B.
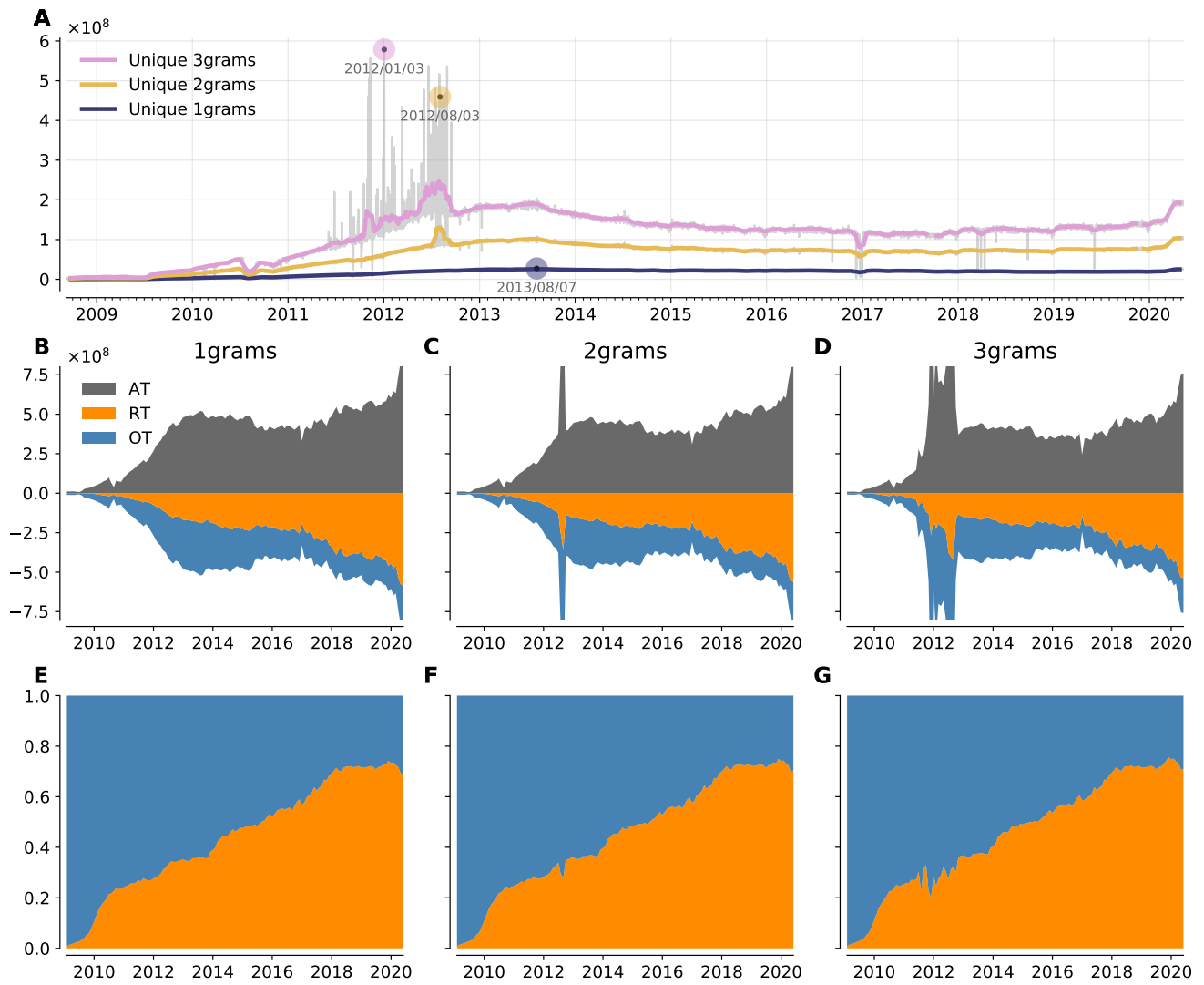
FIG. S7. **Temporal summary statistics**. The grey bars in **(A)** show the daily unique number of n-grams, while the lines show a monthly rolling average for 1-grams (in purple), 2-grams (in orange), and for 3-grams (in pink). In panels **(B–D)**, we show the growth of n-grams in our dataset by each category where n-grams captured from organic tweets (OT) are displayed in blue, retweets RT in green, and all tweets combined in grey. Panels **(E–G)** show the normalized version to illustrate the growth of retweets over time.
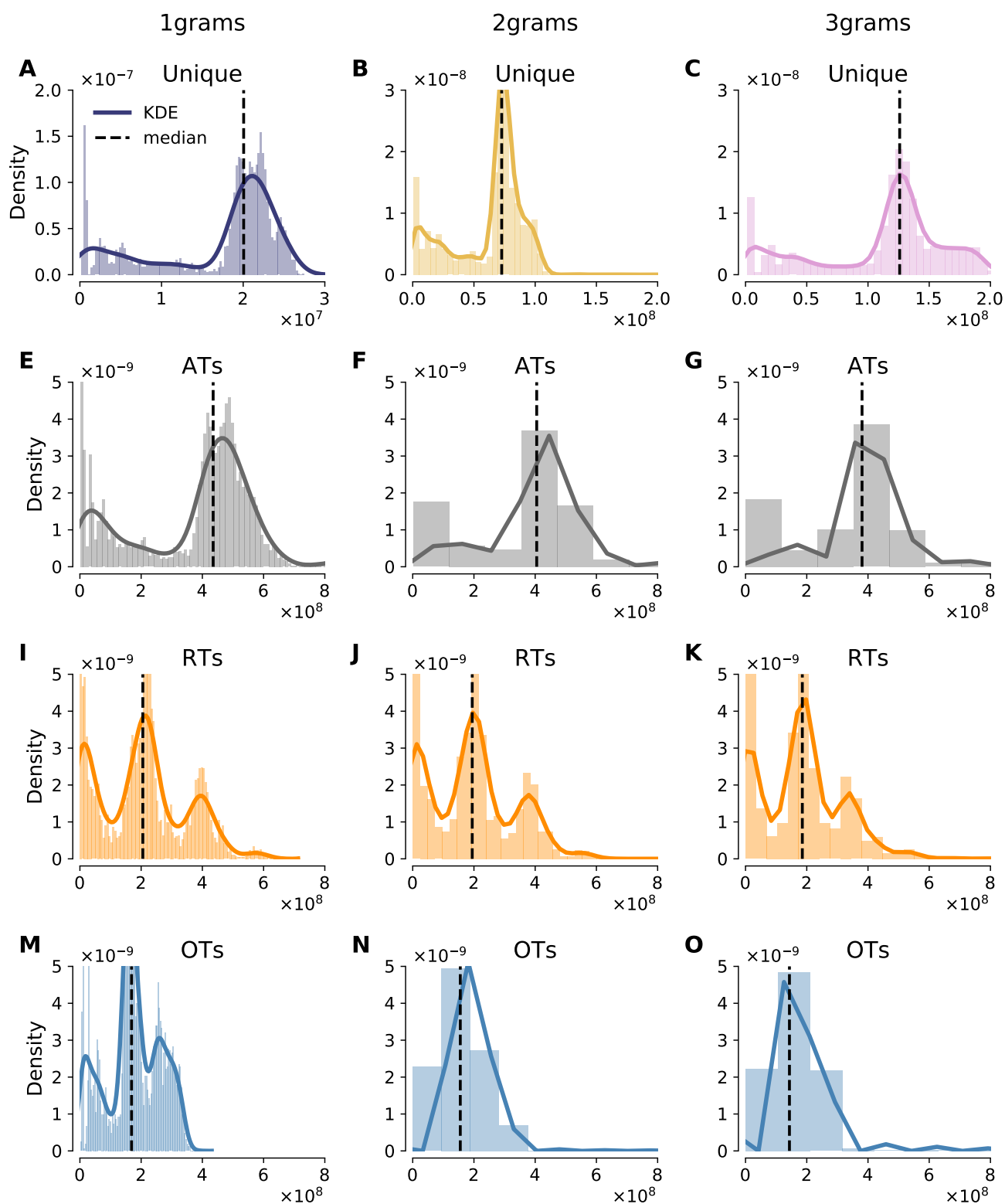
FIG. S8. **Kernel density estimations. A–C.** Distributions of unique 1-, 2-, and 3-grams captured daily throughout the last decade. **E–G.** Actual distributions of *n*-grams occurrences in all tweets. **I–K.** Distributions of *n*-grams parsed from retweets (RT) only. **M–O.** Distributions of *n*-grams parsed from organic tweets (OT) only.