Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/2/4/e1501158/DC1

Supplementary Materials for

The dynamics of information-driven coordination phenomena: A transfer entropy analysis

Javier Borge-Holthoefer, Nicola Perra, Bruno Gonçalves, Sandra González-Bailón, Alex Arenas, Yamir Moreno, Alessandro Vespignani

Published 1 April 2016, *Sci. Adv.* **2**, e1501158 (2016)
DOI: 10.1126/sciadv.1501158

The PDF file includes:

A. Data, Context, and Chronology of the Events Analyzed

We considered five different events: the Spanish 15M protests, the *Outono Brasileiro* (Brazilian autumn) movement, the announcement of the Higgs boson discovery, the release of a Hollywood blockbuster movie (Batman "The Dark Knight Rises"), and the acquisition of Motorola by Google. In this section we report details concerning these events and the associated Twitter data sets.

A.1. The 15M Protests (May 2011).

These protests emerged in Spain in the aftermath of the so-called Arab Spring. A grassroots social movement, later called the *Indignados* ("the outraged"), it emerged from online communication amongst a decentralized network of citizens and civic associations. Online networks (blogs, Facebook, Twitter) were used to spread a call for action for May 15, 2011. The main drivers of the protests were spending cuts and policy reactions to the economic crisis. Massive demonstrations took place on May 15 in several major cities around Spain, many of them resulting in camp sites in main city squares that remained active for weeks. Mainstream media didn't cover the movement until it reached the streets. As a consequence, most communication and broadcasting announcing and discussing the mobilizations took place through online channels. Social media networks (in particular, Twitter) served a crucial role in the coordination of the protests and the management of camp logistics [40,45].

The Twitter data for the Spanish 15M movement were harvested by a startup company (*Cierzo Ltd.*) for a period spanning from April 25 to May 25, 2011. The main demonstrations in Spain took place on May 15th and onwards, thus our analysis covers a brewing period with low activity rates (up to May 15th, day 20 in the Figures) plus an "explosive" phase beyond that date, in which the phenomenon reached general public and was widely covered by traditional mass media, see Figure 1 in the Main Text. Scraps on Twitter servers yielded 581,749 messages.

A.2. Outono Brasileiro Protests (June 2013)

More recently, massive protests filled the streets of several Brazilian cities. The triggering factor was the rising prices of public transportation, but on the background loomed long-standing discontent with inequality, the government economic policies, and the provision of social services. Social media played again an instrumental role in the coordination of large-scale mobilization and the swift diffusion of information: images documenting the often brutal police reaction to the protests boosted mobilization and brought more people to the streets of more cities and municipalities. The protests, often dubbed as *Outono Brasileiro* ("Brazilian autumn"), resulted in Brazilian President Dilma Rouseff announcing, in June 21, measures to improve the management of public transport along with other social services. This prime-time televised address, however, did not placate citizens dissatisfaction, who continued staging protests in subsequent days.

The data set regarding such event has been obtained using the *PowerTrack* tool that provides 100% coverage for a set of specified keywords (see Table S1). For our analysis we considered just the tweets sent in the month of June 2013 (2,670,933 tweets). Indeed, the first large scale protest, often associated with the escalation of the protests, took place on June 17th, with remarkable (though smaller) precursors on the 6th and 13th. As in the case of the Spanish movement we considered a brewing period with low activity rates plus the "explosive" phase beyond the date of the first massive street protest.

AnonymousBrasil	boicot	cacerolada	cacerolazo	huelga
marcha	marchado	marcham	marchamos	marche
march\'e	marcho	Passeata	protesta	protestam
protestar\'as	protestarem	protestarmos	proteste	protestemos
protesten	protesto	protest\'o	concentraci\'on	reforma
greve	rali	manifestaçao	manifestantes	corrup\c c\~ao

Table S1. List of keywords used to find tweets related to the Outono Brasileiro.

A.3. The Higgs Boson Discovery Announcement (July 2012)

In July 4 2012, a team of scientists based at CERN presented results that indicated the existence of a new particle, compatible with the Higgs boson (the existence of which had first been hypothesized in 1964). Mainstream news media covered the discovery after the announcement, but during the days preceding it there were already rumors of its discovery circulating through social media [29]. The messages we analyze were collected using Twitter's publicly available API between July 1st and July 7th using a list of relevant keywords (i.e. LHC, CERN, boson, Higgs). In total, the data set contains 985,590 tweets.

A.4. The Hollywood Movie Release (July 2012)

The Dark Knight Rises is the third installment of the Batman trilogy (started in 2005 with the release of Batman Begins and followed up in 2008 with The Dark Knight). It was premiered in New York on July 16th 2012, and released in several English-speaking countries a few days later. The promotional campaign included so-called viral marketing through social media. The film was nominated to several prestigious awards, and grossed over a billion dollars in the box office.

The data set includes 130,529 tweets between July 6th and July 21st that include the words “batman”, “darkknight” or “darkknightrises”. The tweets are obtained from the Twitter Gardenhose (a 10% random sample of the entire Twitter traffic).

A.5. The Google-Motorola Acquisition (August 2011)

On August 15th 2011, Google announced a relatively unexpected agreement to acquire the mobile company Motorola. The move was a strategic attempt to strengthen Google’s patent portfolio in a context where legal battles over patents is increasingly shaping the mobile industry and the telecommunications environment.

The data set contains 10,890 tweets between August 5th and August 20th, 2011. In order to minimize the noise, we considered just tweets containing both “Google” and “Motorola”. Also in this case, the tweets are obtained from the Twitter Gardenhose.

B. Methods used in the analysis

In this section we detail all the methods that have been used throughout the work: Twitter time series construction and temporal coarse-graining (B.1, B.2); Symbolic Transfer Entropy (STE) measurement (B.3); time scale determination as a maximization of STE (B.4); and central-to-distributed transition characterization (B.5, B.6).

B.1. Data spatial aggregation

With activity information at hand, a possible way to represent information is to assign a time series to individual Twitter users. This however has important drawbacks: activity may be too sparse to build a significant series; it may be rather difficult to detect general, meaningful trends when studying series interaction; finally, one needs to take into account computational costs. Thus, we have chosen to coarse-grain the data from a geographical point of view. Some considerations need to be made at this point:

I. Although the STE does not consider signal volume, the geographical partition has an upper and lower cut-off for the information significance of the social media signal. Generally speaking, if we go down to the scale of a single village or neighborhood, alongside computational issues coming from the dimensionality of the data, driver and driven regions will be too fragmented to provide a clear pattern. On the contrary, if we aggregate the signal on too large regions we will be integrating all partitions into a unique driver area thus making the signal homogeneous and uninformative. This is a long standing issue that affects several other areas ranging from gravity law in transportation science to epidemic modeling in meta-population structure. In all these cases, as in our approach, a best practice is to perform sensitivity analysis for the identification of a meaningful geographic scale.

II. Other alternative partitions are possible: instead of geographical boundaries, we could have used age, religion, gender or other grouping categories, if that information had been available. It is likely that alternative partitions yield even higher STE values, placing a challenging multi-optimization problem. In this work, the quest for an STE-optimal partition is beyond the scope: we chose events for which geography makes sense, i.e. a large corpus of sociological literature has highlighted in the past the role of geography in information diffusion, economic transactions, personal interactions, and civil and political movements [43,47, 54-57]

For the 15M case, geographical information was collected for each user involved in the protests, thereafter tweets were assigned their author's location. Spanish Twitter activity is spatially coarse-grained according to the list of metropolitan areas defined by the European Spatial Planning Observation Network (http://www.espon.eu). This process yields 56 aggregate time series, each corresponding to a geographical area, plus an extra signal which accounts for any activity not included in the previous definition. Thus, the system is made up of N=57 components. Time-stamps have been modified when necessary (Santa Cruz de Tenerife, Orotava and Palmas de Gran Canaria) to a common time frame. The pre-defined metropolitan areas account for over half Spain's total population.

The Brazilian tweets have been instead aggregated at the level of N=97 basins centered around major transportation hubs. These geographical units, that correspond to census areas surrounding large cities, have been defined aggregating population cells of 15×15 minutes of arc (approximately a rectangle of 25×25 kms along the Equator), from the "Gridded Population of the World" and the "Global Urban-Rural Mapping" projects [41,42], to the closest airport that satisfies the following two conditions: (i) Each cell is assigned to the closest airport within the same country; and (ii) the distance between the

airport and the cell cannot be longer than 200km. This cutoff naturally emerges from the distribution of distances between cells and closest airports. See refs. [35,39] for details. Moreover, having access to 100% of the entire signal on Twitter associated with at least one word listed in Table S1, we considered just tweets with live GPS coordinates.

Tweets around the Higgs boson discovery were aggregated at the country level. The original data set contained tweets from over 200 countries, but these have been thresholded to retain only those countries with more than 500 tweets over the topic of interest, for the whole week. This entails that only N=61 countries are present in the analysis in the main text. The details regarding the location technique can be found in [29].

The remaining two data sets have been aggregated in N=52 areas –50 U.S. states, plus Washington D.C and Puerto Rico. In particular, the geographical information of tweets in this case has been gathered either from live GPS locations, or mining the so called "self-reported location". In general this field is filled freely by the users that can report their location at different levels. i.e. NYC, California, CA, USA etc.. Some fraction of the reported locations are jokes, i.e. Moon, Mars, behind you, etc. We parsed these fields trying to match a country, state, or city name. In fig. S1 we report the flow chart of the algorithm used.

Interestingly, the method is able to find a match of 40-50% the total number of tweets. It must be noted that identifying the location of a user is a hard problem, so some assumptions must be made, and with each assumption noise is introduced in the data. Even the simplest assumption possible, that a user is where her GPS coordinates say she is has been shown to not always be true [58]. It's also clear that the user is not always where she claims to be on her self-reported location field, as this field is infrequently updated even when users travel. On the other hand, an argument can be made that the self-reported location field (even if it does not correspond to the instantaneous location of a user) corresponds to the city or region that the user identifies with, which is relevant for our case.

In summary, while we acknowledge that the two types of location may correspond to two different things (current location *vs.* "home" location), these are precisely the two types of location that are relevant to the aim of identifying which users are invested in the political activities occurring in each city, as (in general) we are directly affected both by what is happening around us and by recent developments back home while away. If our goal is to understand how developments in one city are driving what is happening in another city, we must take the most significant location information available. We believe that both a GPS coordinate and self reported location are relevant.

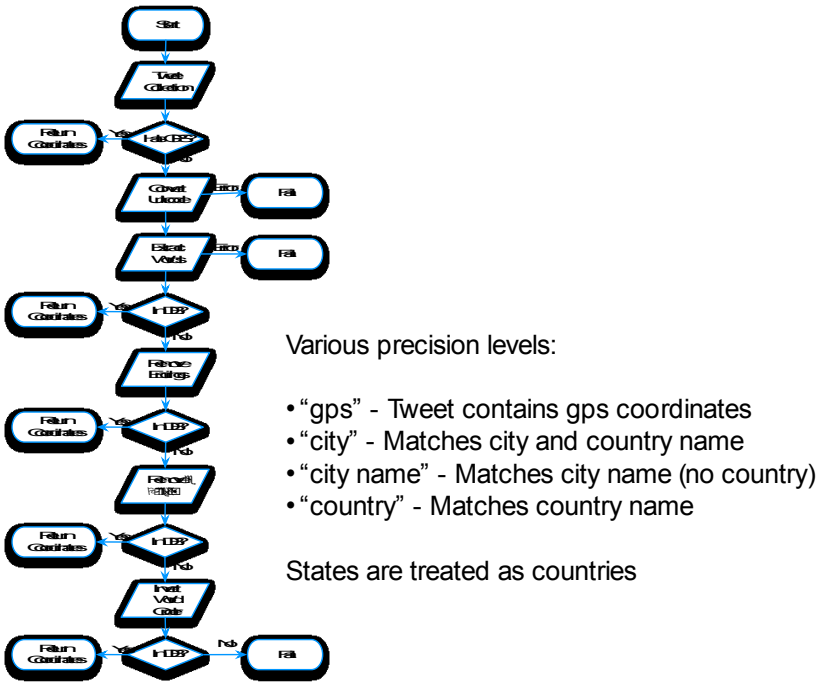

fig. S1. Schematic representation of the algorithm used to gather geographical coordinates of the Hollywood movie release and the Google-Motorola acquisition data sets. In the figure DB stands for database. We used the "GeoNames" database [44].

In subsection C.3 (below) we show the results obtained considering different geographical aggregations. Namely, we also considered N=16 Spanish communities (one of them was left out because no messages were collected from there), N=27 Brazilian states, and N=9 regions in the USA defined by the American Census Bureau [36].

B.2. Data temporal aggregation

The definition of the temporal aggregation of Twitter data is particularly important in our approach. Indeed, we want to determine the characteristic time scale at which the driving between series is most evident. The data comes with temporal resolution down to a second. However, such level of resolution is excessive to detect dynamical trends among series. We considered different sampling rates Δt spanning from 1 to 120 minutes. Although arbitrary, this range of temporal aggregations account for the fluidity of Twitter's discussion as well as the limited attention time span of users. In subsection B.4 we discuss the ideas that allow us to determine the optimal temporal aggregation scheme.

B.3. Symbolic Transfer Entropy

Closely related to other measures, such as mutual information [53], Granger causality [46] and transfer entropy [23], Symbolic Transfer Entropy (STE) [21] provides a solid method to detect and quantify the strength and direction of couplings between components of dynamical systems. The symbolic approach, on the other hand, links STE to order patterns and symbolic dynamics [22,49] as a means to successfully analyze time series which may be noisy, short and/or non-stationary.

Once spatial and temporal aggregation schemes are fixed, we proceed to measure STE as a way to quantify the coupling among series. Note that such series span long times, L, of several days or even a month. Also, activity during these days is changing due to offline events happening outside the Twittersphere. Thus, STE is not measured over time series taken as a whole, but over sliding windows of length $\omega << L$ (which is indeed a standard way to proceed in other relevant fields such as neuroscience). To obtain a finer analysis, these windows advance at a slow pace of only 30 minutes. In practice, this means that the first window spans the interval $[0,\omega]$; the second one $[30, \omega + 30]$ (in minutes), and so on. Window width ω, admittedly, is the first parameter that will affect the measurements output, and we will discuss its effects later (subsection C.1 below).

Given a window of width ω, the resulting series are transformed into symbol sequences as described in [21], for which an embedding dimension $3 \leq m \leq 7$ [22] must be chosen (see also subsection C.2 for further details). Let us consider a simple example of how this works. Imagine we have a signal

$$x = \{120, 74, 203, 167, 92, 148, 174, 47\}$$

(let us ignore sliding windows by now). We shall transform this series into symbol series. For simplicity, let us suppose that the embedding dimension m=3. This quantity determines the amount of symbols that can possibly exist, which is $m! = 6$ in our case. See fig. S2 as an illustration of the possible symbols that can be obtained.

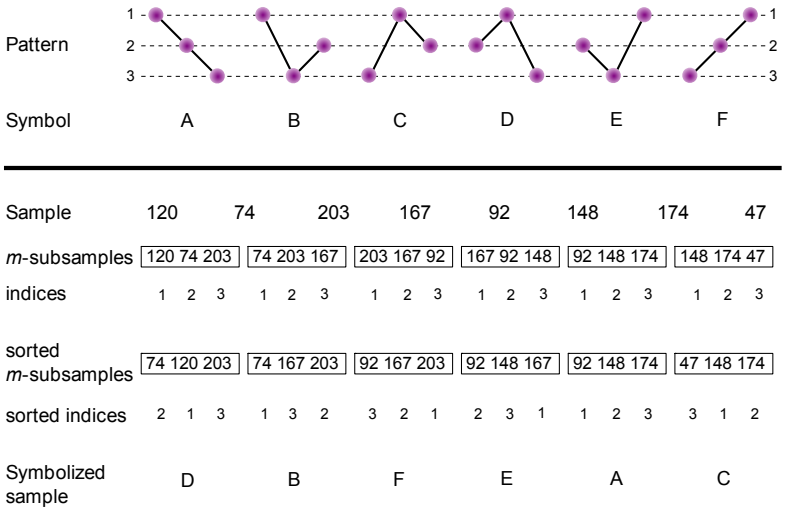

fig. S2. Sample order pattern for $m = 3$. If we neglect ties, the number of possible patterns is 3! (see [21] to check how ties are dealt with).}

The first step to transform x into symbol sequences is to sort their subchains of length m in increasing order. So, we take the first three elements of x and sort them, which leaves us with {74, 120, 203}. We have kept track of these values' indices, such that the sequence now looks like {2, 1, 3}. According to fig. S2 (top), this first subchain maps to the symbol D.

From this scheme, we just need to advance one value at a time: the next subchain to consider is {74, 203, 167}. Its sorted version is {74, 167, 203}, which corresponds to {1, 3, 2}, and maps to B. The whole process for the x signal looks like fig. S2 (center), and their sorted indices lead to fig. S2 (bottom), rendering a symbol sequence $\hat{x} = \{D, B, F, E, A, C\}$. With a similar procedure, other series y are transformed into $\hat{y}$. Given these symbol sequences $\hat{x}$ and $\hat{y}$, STE between a pair of signals (x, y) is defined as in Equation (1) of the Main Text.

A few facts need to be highlighted at this point:

I. A signal with an original length of n points is reduced, through symbolization, to a new string with n-m+1 symbols.

II. A way to interpret the meaning of m is to think of it as the amount of "expressiveness" it allows to the original series. That is, if m is low, a rich signal (one with many changes in it) is reduced to a small amount of possible symbols. This is of utmost importance to understand why we have chosen a relatively high m to work with (see subsection C.2).

III. $m = 7$ is not a *theoretical* upper limit, but rather a *practical* one: as the alphabet size scales as $m!$ the computational costs beyond that value become unfeasible (we follow ref. 22 in this aspect). Besides, there exist ways to quantify how suitable a certain embedding dimension is, i.e. false nearest neighbors (fig. S7 below). In most cases, $m > 5$ does not imply a better projection of the data in the symbol space.

IV. All measurements in the present work have been performed using $\delta = 1$. This implies that we are measuring the capacity of a signal to predict the *immediate* future of another signal, i.e. just one symbol ahead.

Finally, the following figure summarizes how STE has been measured in all data sets throughout the main text and SM, specifying the meaning of each parameter (s, m, Δt, etc.) and the range of values that have been used in each case.

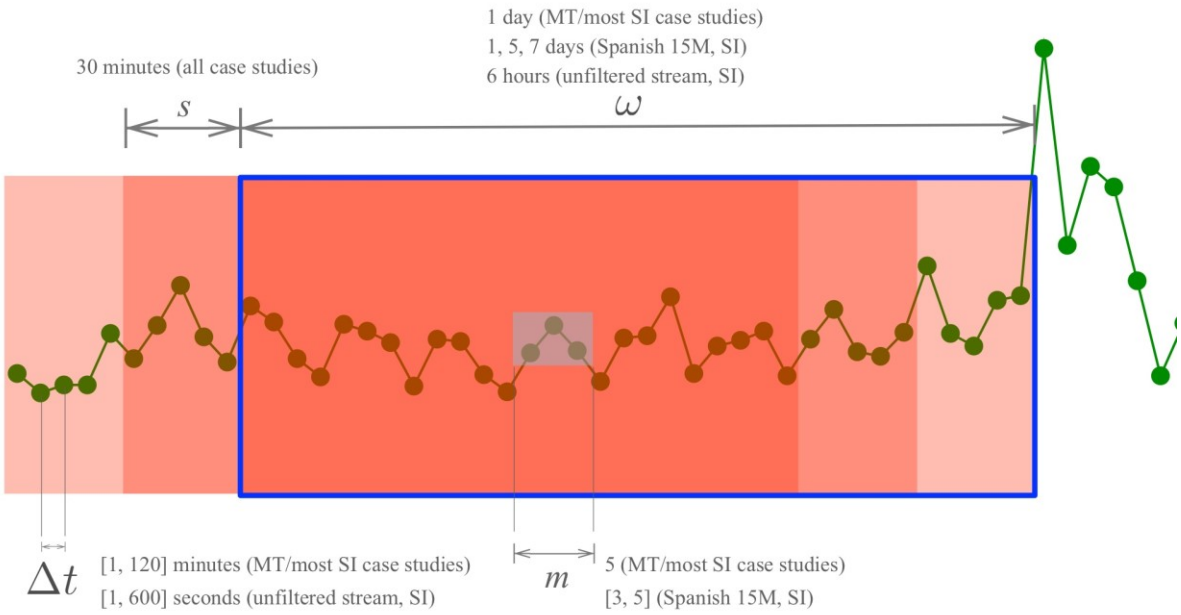

fig. S3. Schematic view of the sliding window scheme. Each snapshot of width ω encompasses a sub-chain of ω / Δ*t* points. This sub-chain is symbolized, taking *m* time series ticks to construct a single symbol. Once STE has been measure for a certain window, the process moves on, selecting a new window at s distance in time from the previous one. A large ω – *s* overlap guarantees that the evolution of the system is tracked smoothly (i.e. it is convenient to choose *s* << ω).

B.4. Defining the characteristic time scale of the events

Given a certain data set, we do not have any prior knowledge to define the correct timescale Δ*t* at which time series should be aggregated. We do know, however, that activity around civil protests (and in general around any event that involves collective action) are typically far from being stationary or periodic which intuitively points at the fact that there might not exist a single time scale for the whole data set.

Let's consider for now a fixed temporal resolution Δ*t* and bin the Twitter activity such that the first point in the time series will contain any activity that happened between [0, Δ*t*); the second point will contain data from [Δ*t*, 2Δ*t*), and so on. For example, a 30 day data set, sampled at Δ*t* = 60 minutes will render a set of time series of length $L = 30 \times 24 = 720$ points.

As mentioned above, the STE is measured using signals in windows of width ω, spanning from [*t* − ω,*t*). In order to define the optimal Δ*t* we evaluate which temporal resolution provides the best possible information flow among units, i.e. the optimal τ = Δ*t* is the one containing more transfer entropy $s_{\Delta t} = \sum_{x,y} T_{xy}^{\Delta t}$. Operatively, the best timescale is defined as $\{\Delta t : \max_i(s_{\Delta t}^i)\}$. We considered as possible candidates all values Δ*t* from 1 to 120 minutes, with increases of 5 to 15 minutes depending on the data set.

It is important to notice that the maximization of STE at each sliding window may result in different optimal Δt: Our methodological proposal infers from the data the time scale at which events are best described, as illustrated in Figure 2 in the main text.

B.5. Defining the information flow of the events

One may further scrutinize the temporal evolution of the amount of STE each subunit displays. Instead of studying pairwise information flows, in this case we focus on whether a geographical unit is on average driving others or is driven by others at each time step. Within each window of width ω at time t, we calculate the values of the net flow matrix for that window, or normalized directionality index (di), for each geographical unit x at each time step defined as:

$$x_{di} = \begin{cases} \frac{\sum_y T^S_{xy}}{\max_x \sum_y T^S_{xy}} & \text{if } \sum_y T^S_{xy} > 0 \\ \frac{\sum_y T^S_{xy}}{\min_x \sum_y T^S_{xy}} & \text{if } \sum_y T^S_{xy} < 0 \end{cases}$$

Thanks to the normalization $-1 \leq x_{id} \leq 1$ the largest value is associated to the geographical unit exercising the largest driving force to other units. Vice versa, the smallest value is associated to the geographical unit subject to the largest driving forces from other nodes. The result of this measurement --and the corresponding analysis-- can be found in Figure 3 in the main text. Note that each point in the panels of that figure condenses the results obtained for a window integrating information from the *past*, i.e. activity within $[t - \omega, t)$.

B.6. Thresholded versus raw "dominance matrices" $T^\dagger$

Figure 5 in the main text represents θ as measured from thresholded matrices (i.e. matrices in which the lowest Transfer Entropy values have been removed). In this section we discuss fig. S4, where the order parameter θ is represented both for thresholded and raw dominance matrices.

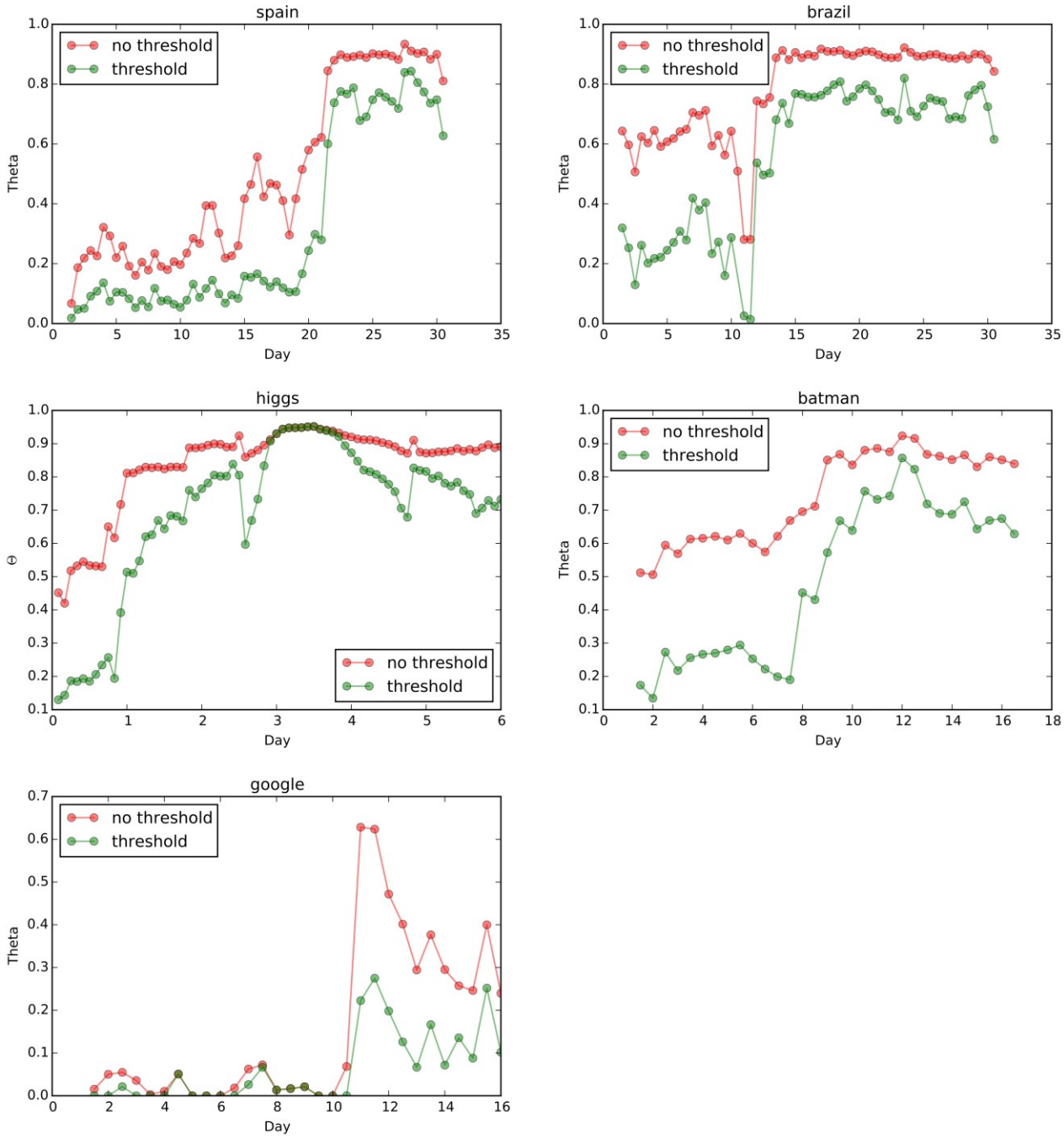

fig. S4. Evolution of the order parameter θ for thresholded (green) and raw (red) $T^{\dagger}$ matrices. Despite a natural increase of the noise in \theta, the interpretation is qualitatively robust.

C. Sensibility analysis of the parametrization

Results in the main text have been obtained with 1-day (i.e., ω = 1) long sliding windows, and embedding dimension m = 5, and using particular geographical units (metro areas in Spain, basins in Brazil, and states in USA). In this section, we present the results of various sensitivity analyses testing how each parameter influences the results.

C.1. The role of the sliding windows width ω

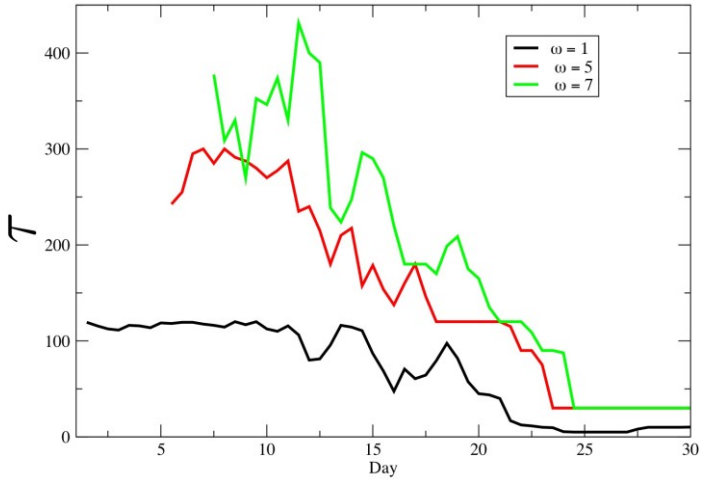

fig. S5. Dependence of τ with he sliding window size ω considering the Spanish 15M protest. Black, red and green lines describe ω = 1,5,7 respectively. Note that, exceptionally, for ω = 5 and ω = 7 days we have considered large Δ*t*, up to 480 minutes (very slow time scales).

Intuitively, if ω is set to a large value the capacity of the method to anticipate events will be reduced, because information emitted long before the present time is affecting the calculations. To check how larger ω's blur the results, we have reproduced the observations in the main text for ω = 5, and ω = 7 given a fixed $m = 5$ considering the 15M data set. Results are offered in fig. S5. It is clear that different window widths behave in the same way as the original one (1 day). Nevertheless, shorter ω yields a more abrupt transition close to the critical event. This observation is more evident studying the behavior of normalized directionality index. In fig. S6 we show this quantity at each time step for the case of the 15M protests in Spain. Similarly to Figure 3 in the main text the size of each bubble is proportional to the logarithm of the activity on twitter. We notice that for ω = 7 and ω = 5 the system shows a change in the driving dynamics clearly before the 15th of May (red strip). Instead, for ω = 1 the transition form a scenario in which the large metropolitan areas are the major driving forces to a more homogenous and delocalized scenario happens during the unfolding of 15th of May.

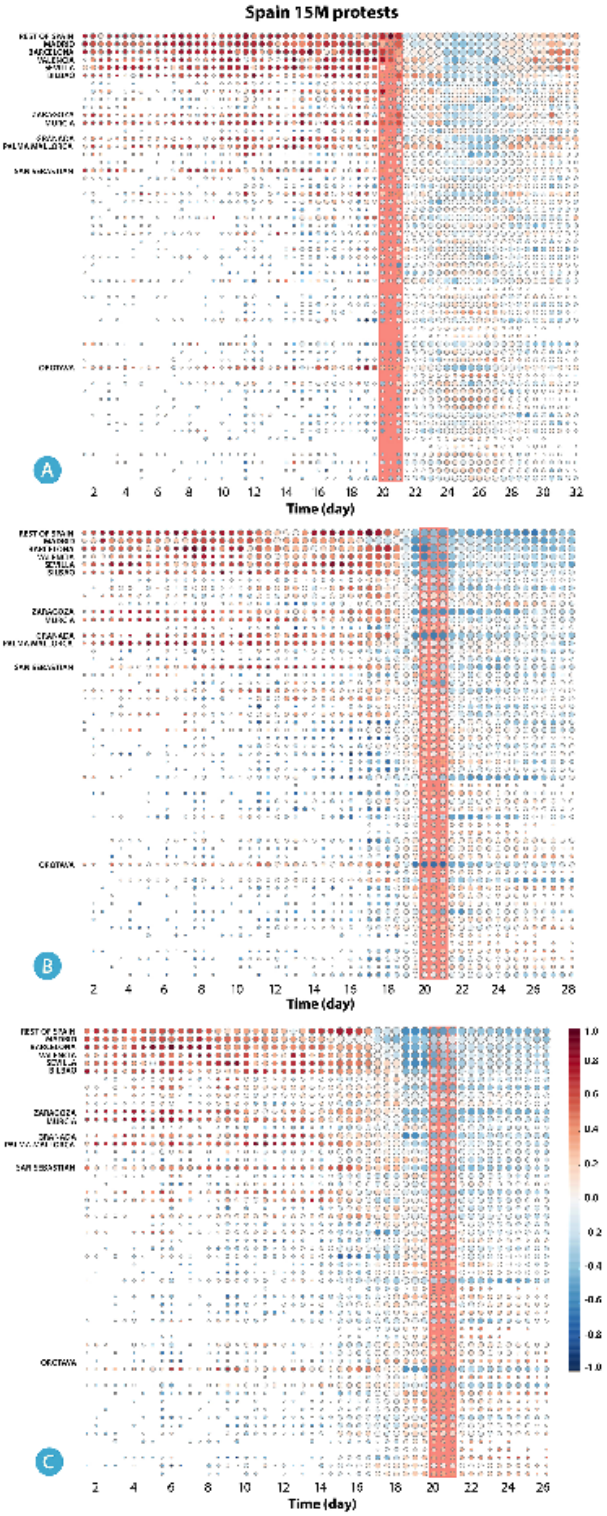

fig. S6. Normalized directionality index for each geographical unit in the 15M data set for different ω. In panel A) we considered ω = 1, in panel B) ω = 5 and in panel C) ω = 7. For all the cases we set $m = 5$. The red strip indicate the 15th of May}

C.2. The role of the embedding dimension m

The embedding dimension m determines how the information in the original time series will be transformed into symbols. The larger m, the larger is the collection of symbols onto which the values are mapped. Since the size of symbols grows like $m!$, it is clear that complex time series demand higher m for a faithful mapping (i.e. one that collects the original complexity). On the other hand, overestimating m adds unnecessary computational costs, because the final result won't change qualitatively. We address the problem of finding the minimal sufficient embedding dimension m, using the approach, called the false nearest neighbor method, proposed by Kennel *et al.* [48].

In practical terms the minimal value of m is found studying the behavior of the nodes encoding the large majority of information. In the case of the 15M protests in Spain this corresponds to Madrid, which was a key spot for the grassroots movements. The minimal value of m is sufficient to disentangle the signal from the dominant node and any other time series in the corresponding data set will need the same or smaller m to be faithfully mapped. Figure S7 reflects these calculations and the strong conclusion is that for any $m \geq 5$ symbolization will have captured the original topology of the real data. Thus, all through the main text, and also in this document, results are reported for $m = 5$, unless indicated otherwise.

In order to further study the effects of m in fig. S8 we plot the behavior of the normalized directionality index for $m = 3, 4, 5$. As it is clear from the plot $m = 5$ (panel C), is able to capture the transition from asymmetric to symmetric scenario in more detail.

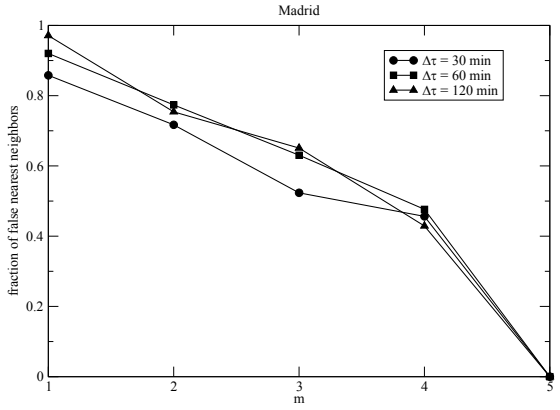

fig. S7. Fraction of false nearest neighbors as a function of m for the Spanish data set and the Madrid time series.

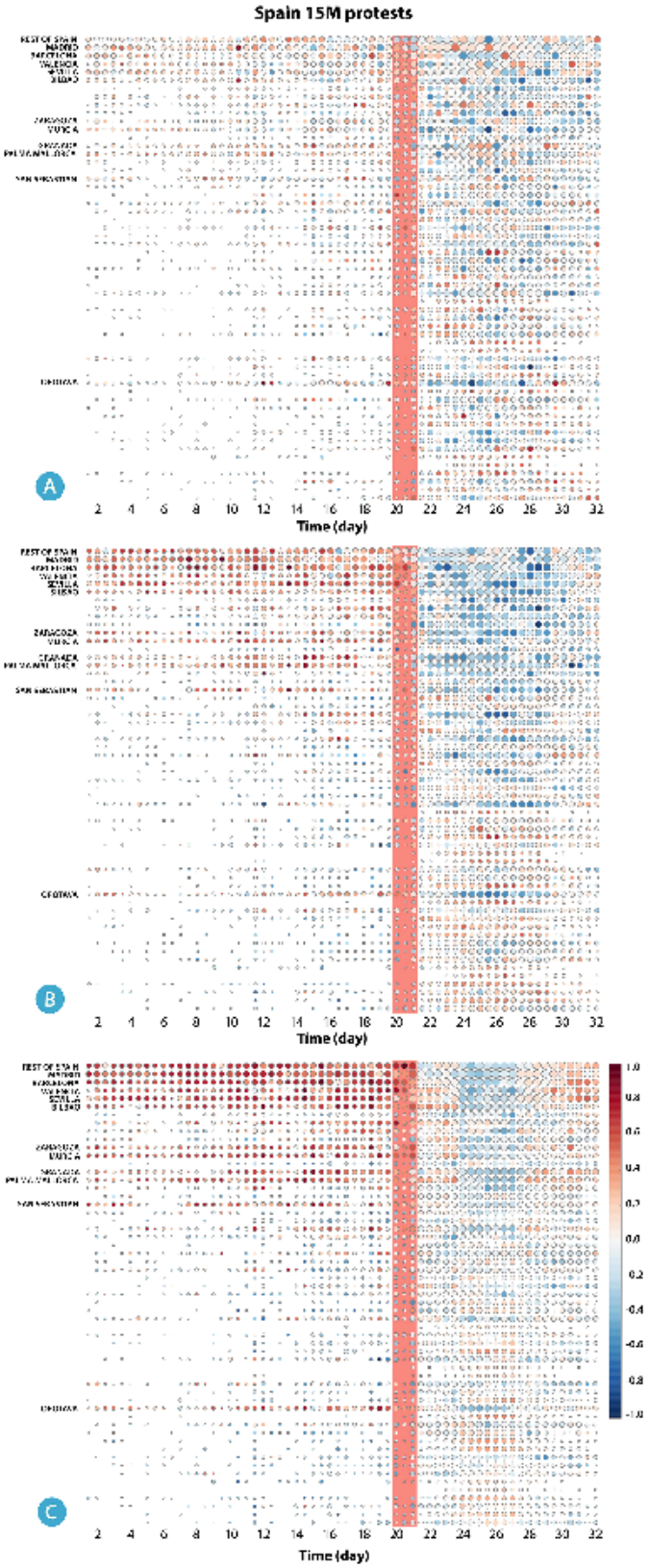

fig. S8. Normalized directionality index for each geographical unit in the 15M data set for different m. In panel A) we considered $m = 3$, in panel B) $m = 4$ and in panel C) $m = 5$. For all the cases we set $\omega = 1$. The red strip indicate the 15th of May.

C.3. Sensibility analysis of the partition

As mentioned above, the Twitter signal could be partitioned in many ways –and we have chosen a geographical approach. In fact, an enormous range of settings are possible: from a simple bipartition of the activity stream to a complete breakdown where a single user is matched to a time series. It is fair then to state up-front that our decision is an arbitrary one, driven by the empirical intuition that geography matters in the emergence of collective action efforts.

Once the geographical partition is adopted, many options are still available: spatial aggregation could be done at the neighborhood, city or county levels (for a finer resolution), or considering a coarser partition. For each data set in the main text we have repeated the analysis for coarser geographical divisions. In the case of Spain's 15M movement, we have moved from the metropolitan areas to the autonomous community level (the Spanish 17 autonomous communities can be regarded as states, i.e. political entities at the regional level [38]). Data from Brazil have been binned in 27 states [37], in contrast with 97 basins in the main text. Finally, US data has been aggregated up to the "divisions" level (9 supra-state areas, as defined by the US Census Bureau [36]).

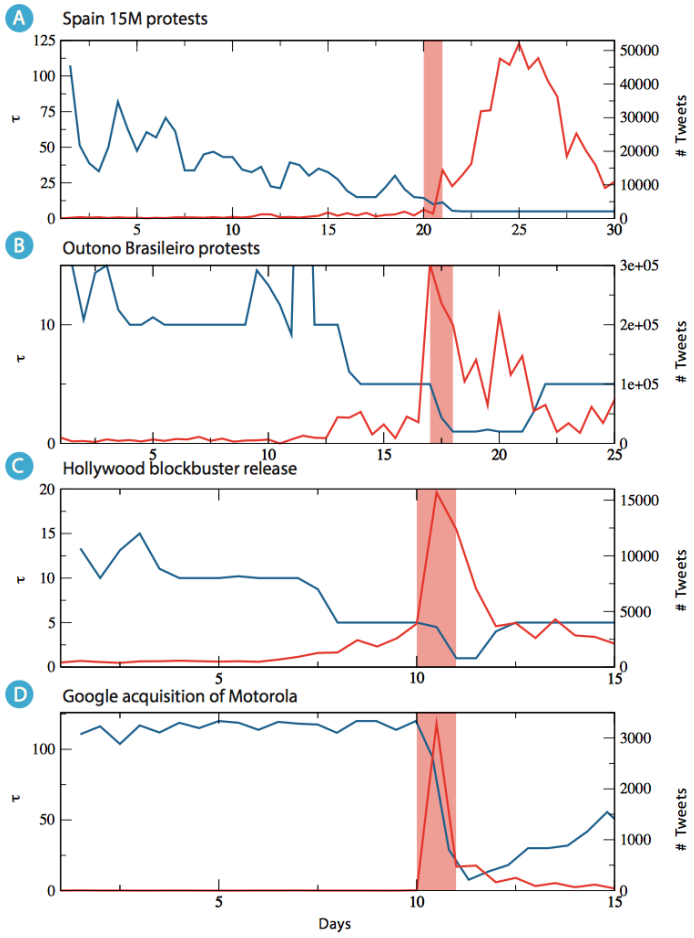

fig. S9. In light blue we plot the characteristic time scale of STE (τ) for the **(A)** 15M protests, **(B)** Outono Brasileiro movements **(C)** release of a Hollywood blockbuster **(D)** the acquisition of Motorola by Google. We considered different geographical aggregations alternative to the ones displayed in

Figure 2 in the main text. The red lines show the activity in Twitter for each data set, and the red strip indicate the day of the main collective event.

Results for these alternative data partitioning can be seen in figs. S9, S10 and S11. Regarding the evolution of the time scales (fig. S9), we observe that the behavior qualitatively resembles the original one in the main text. Time-scales appear to drop at even earlier stages (see fig. S9A); indeed, larger aggregation implies that certain subsystems begin to have a noticeable dynamic role earlier (where previously scattered signal remained insignificant overall). Note however that earlier changes in the time scale come at the cost of a loss in precision, regarding which specific areas are true drivers of the activity (see fig. S11).

A similar result is obtained for the information flow balance in fig. S10, where the occurrence of protests and demonstrations (15M and *Outono Brasileiro*) marks a change in the dominant pattern; the same can be said for the Google-Motorola case.

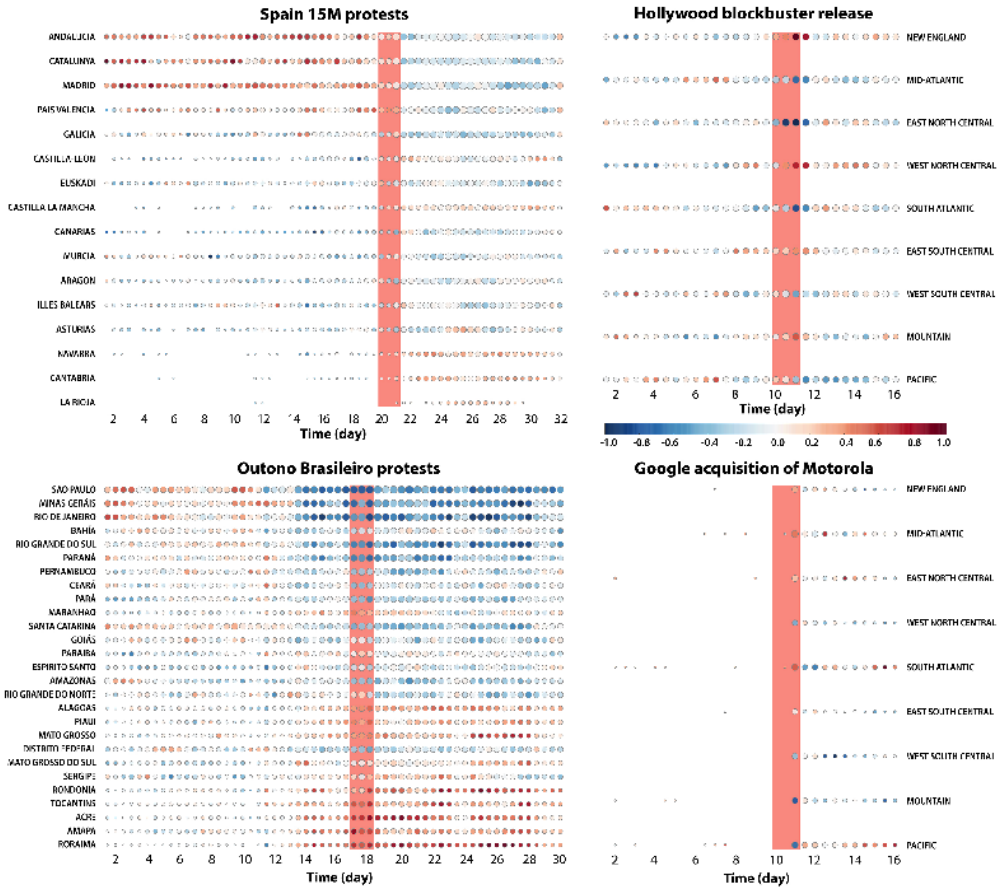

fig. S10. Normalized directionality index for the four data sets aggregated at different geographical levels with respect to those used in the main text. Note that differences with Fig. 3 on the main text are simply due to the new coarsening, and that even in this case the global pattern occurs at similar times and they are just more pronounced.

However, some differences appear in the $\theta - t$ plot (fig. S11). To start with, the Brazilian data set and the Batman event deliver dense T_{xy} matrices, i.e. its sorted counterpart T' has many below-diagonal elements even at early times, indicating that no (or little) transition takes place: the system is decentralized from the very beginning. These differences demand some explanation.

First, it must be highlighted that our tip-over rationale (see Main Text) is valid regardless the apparent contradiction: our claims are concerned with how the values in the T_{xy} matrix are distributed, and as such it is an abstraction of what such matrix represents (be it cities, states or individuals, for that matter). Then, the apparent contradiction simply points at the fact that the lens through which we analyse the events *does* matter. Taking it to the extreme, a bipartition of the data, with two time series accounting for half of the activity each, would easily yield a fully symmetrical T_{xy} matrix; in the opposite situation, a system comprising each user individually would render an (almost) empty matrix, given the fact that most people is not showing activity most of the time.

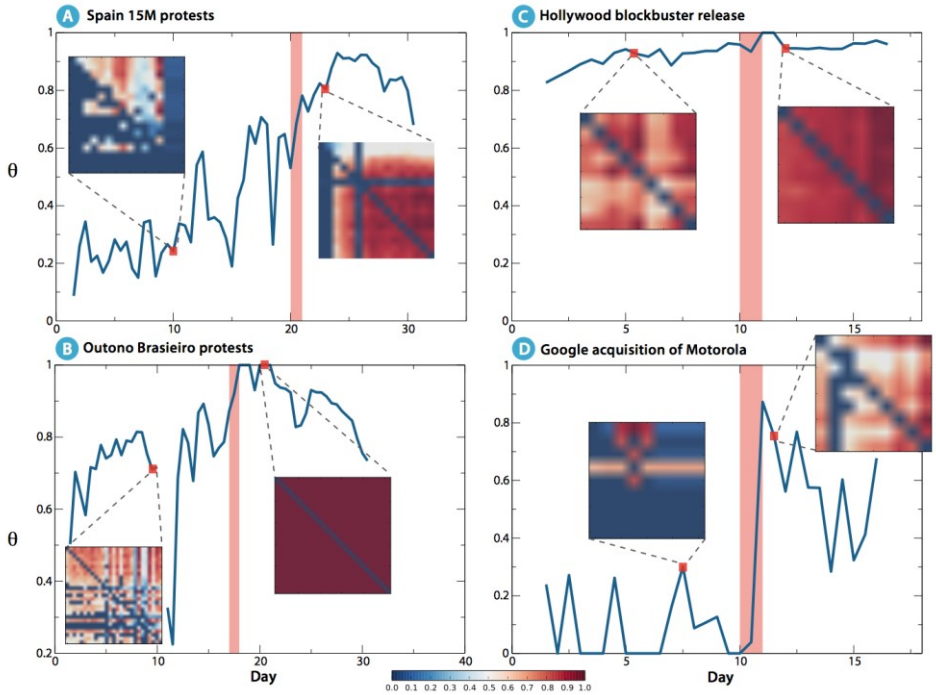

fig. S11. Behavior of θ as a function of time for different geographical aggregations. For each data set two matrices T' are plotted considering a time before and one after the main event (signaled with a red vertical red bar). A blurred transition can still be observed for events A and B. Note a point missing in the Brazilian data set due to a data blackout between days 10 and 11.

All in all, these results suggest that our proposal opens up exciting research questions: for instance, at which level of resolution should the system be observed to extract an optimal analysis out of it? We must keep in mind that other relevant events in Twitter do not have a geographical component; groups may be defined by religious beliefs, age, gender issues, or other demographic attributes. It remains beyond the scope of this work to determine how to obtain optimal partitions that will render the correct conclusions. For the time being, we rely on commonsensical, predefined –rather than optimally detected– entities (geographical, in this case) to make a case of our methods and rationale.

D. Validation of results (I): Time series randomization

In this section we validate our framework studying its performance on data surrogates, i.e. statistical ensembles of randomized data. In other words, we apply our approach to a set of data that by construction do not contain the temporal correlations we find in real data sets. This step is crucial to prove that our observations capture genuine features of real collective events.

D.1. Statistical randomized surrogates of original data

In order to validate our results, we need to make sure that our analysis and conclusions are not mere artifacts which would arise in any case. To provide a reasonable baseline, we need to build randomized counterparts of the data and then analyze it just as we did for the actual case. We consider two methods to obtain data random surrogates: amplitude adjusted Fourier transform surrogates and constrained randomization surrogates, with an extensive use of the TISEAN software [50].

We present the results for the randomization of two data sets, with qualitatively similar insights for the other data sets.

D.2. Amplitude Adjusted Fourier Transform (AAFT) surrogates

A first, robust step to provide a suitable null model is to generate randomized data sets which ensure that certain features of the original data will be preserved. In particular, we generate AAFT surrogates as proposed in [51], who established an algorithm to provide surrogate data sets containing random numbers with a given sample power spectrum and a given distribution of values.

Under these constraints, we obtained 50 randomized versions of the 15M data, which were then analyzed in the same way as the original data (see Main Text and previous sections). The averaged results from such analysis are offered in fig. S12. Clearly, the original patterns are completely blurred and just a single characteristic time scale can be observed. Furthermore, our approach do not capture any change in the characteristic time scale as correlations and driving between different units have been artificially eliminated in the data.

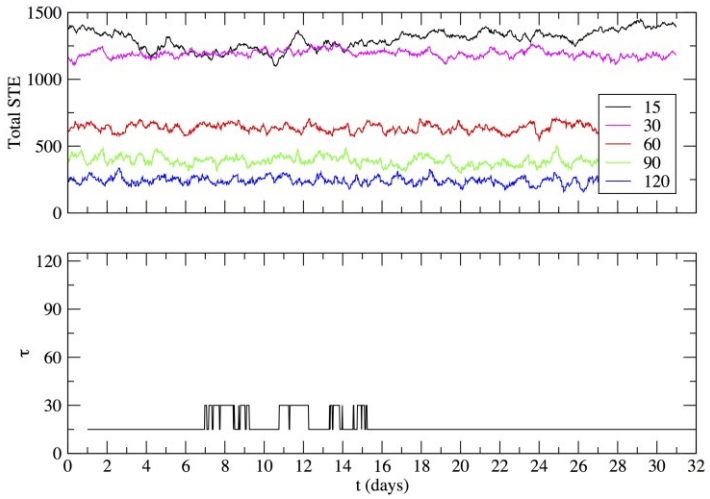

fig. S12. Average total amount of STE for some Δt (top panel) and time scale profile τ (bottom panel) for 15M amplitude adjusted Fourier transform surrogates (50 randomizations).

Mirroring our analysis of real data, we intend to see whether some trace of the original transition is kept for this newly obtained random version of the data. To do so, we also exploit 50 randomizations of the original data set, for which we can extract average surrogate snapshots (i.e. the state of the system at a given *t* day). Just as in Figure 5 of the main text, fig. S13 shows two sorted (ranked) $T^{\dagger}$ matrices, corresponding to two different moments (before and after the main event) for the statistical randomized surrogates of original data. It is interesting to notice that as any localized abrupt change in the time-scale is washed out (fig. S12), also any sort of systemic transition is missing.

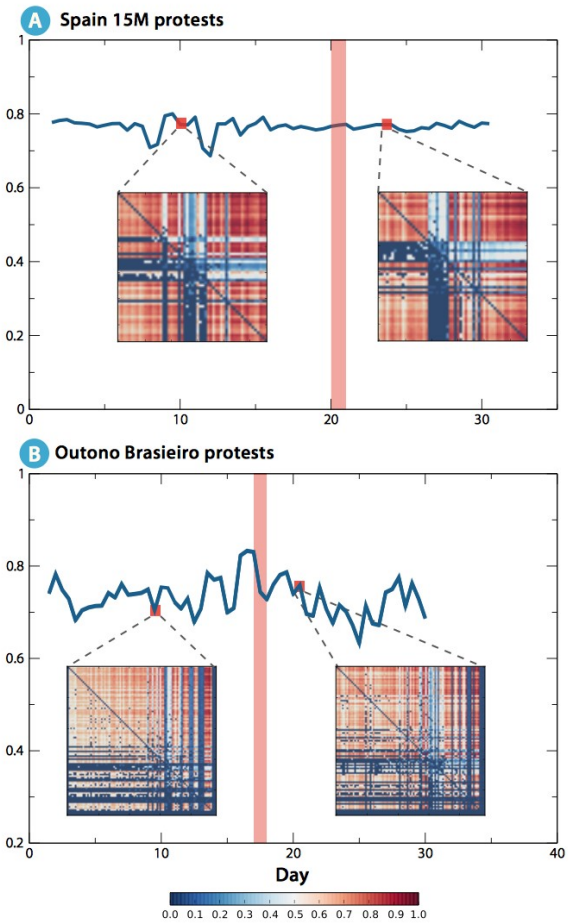

fig. S13. Behavior of θ as a function of time for the statistical randomized surrogates of the 15M **(A)** and Brazilian data set **(B)**. For each data set two matrices $T^{\dagger}$ are plotted considering a time before and one after the main event (signaled with a red vertical red bar). As expected in this context, no significant transition is observed in neither cases.

D.3. Constrained randomization surrogates

Beyond a randomization scheme that guarantees given power spectrum and distribution of values, one might want to generate surrogates which are further constrained. This can be achieved if we demand randomized data sets to preserve as well a given non-periodic autocorrelation function (ACF). To this end, Schreiber [51] developed a method of constrained randomization of time series data which seeks to meet the given constraints through minimization of a cost function, among all possible permutations, by the method of simulated annealing.

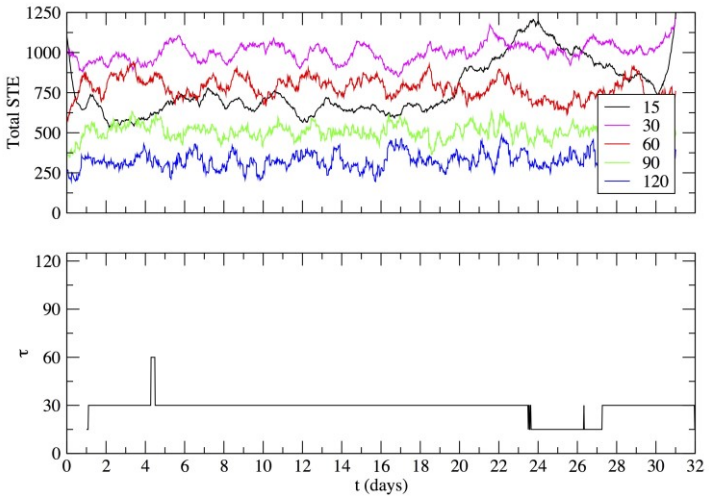

fig. S14. Average total amount of STE for some Δt (top panel) and time scale profile τ (bottom panel) for 15M constrained surrogates (20 randomizations).

In fig. S14 the averaged results for the analysis of the constrained surrogate data can be checked. Twenty randomizations for the Spanish data set were obtained. Even with the additional constraints (if compared with AAFT randomizations, see previous section), hardly any resemblance with the original patterns can be observed. It must be noted that constrained randomizations are time and CPU-consuming, due to the additional restrictions regarding ACF.

E. Validation of results (II): Unfiltered Twitter stream

To further test the contributions of the work –a methodological framework to reveal the evolving time scales of collective events, and its inherent order-disorder transitions– we entertained the idea of analyzing "less successful events". However, such an idea is difficult to implement –if events are less successful, most likely it is not possible to know which are the appropriate words to filter the Twitter stream. Therefore, we propose the opposite: we have collected an *unfiltered* sample of geocoded tweets from all around the world, and analyzed its development for some time.

Some facts:

- Only 2 days have been analyzed (enough to grasp any pattern, if it exists, given that we are not monitoring any specific event). This roughly corresponds to ~3M tweets with precise geolocation.

- We have monitored the stream with window width $\omega = 6$ hours (note this is a smaller window than for our previous analyses, where $\omega = 24$h).

- Each 6 hour window "advances" 30 minutes to analyze the next snapshot (this has not changed from our previous analyses)

- The embedding dimension has been kept at $m = 5$.

- Original Twitter data contains tweets from 230 countries. Tweets from US are most frequent, accounting for ⅓ of the total (~1M); Niue is least frequent (1 tweet). To keep computational costs at a reasonable level, we have arbitrarily thresholded the list of countries. We have set that threshold at 5000 tweets (in the whole period), retaining only 42 countries.

- The set of Δt values that have been used is: 1, 5, 10, 15, 30, 45, 60, 120, 180, 300 and 600 seconds (in brief: a good sample between 1 sec and 10 min). Note that in our case studies we were in the order of minutes, not seconds (in unfiltered scenarios, we foresee the time scales to be very fast).

Results:

The resulting STE-optimal time scales (τ) analysis yields an almost constant value (45 sec) with a few ticks in which it is 30 sec, see fig. S15 below. This is indeed expected. Additionally, this finding strengthens the idea that large volumes are sufficient to derive fast time scales. The fact that we find fast time scales in the *absence* of large volumes enhances the value of the proposed tools.

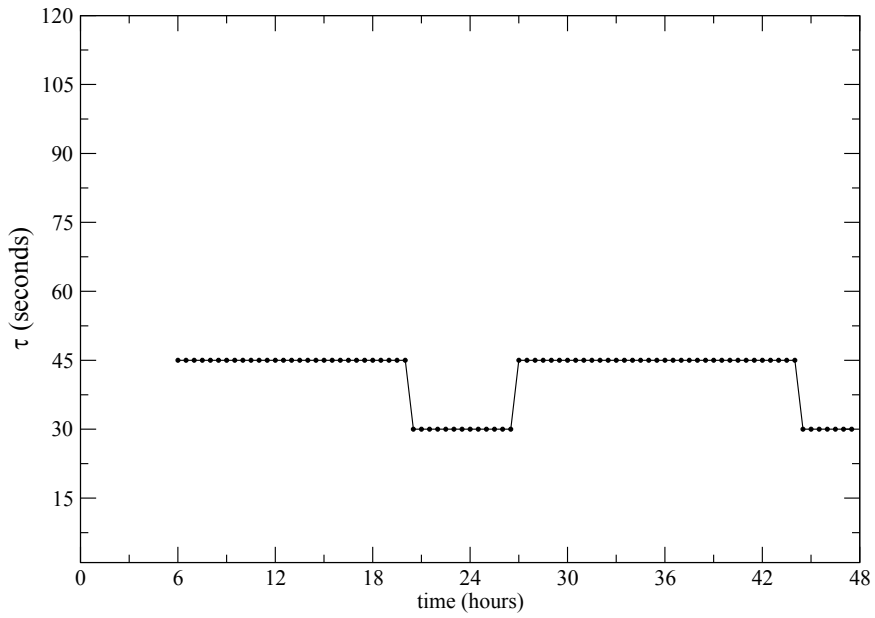

fig. S15. Evolution of τ as a function of time. In an unfiltered scenario, there are no distinguishable time scales to characterize the system's development. For the sake of clarity we represent here up to 120 sec on the y-axis, despite calculations were made up to 600 seconds.

In fig. S16 below we see two thresholded $T^{\dagger}$ matrices corresponding to different moments in the unfiltered data set: the very first one (from 0 to 6h), and one from the final part of the studied period (from 36h to 42h). These are good representatives of any of the snapshots, in which clearly no pattern is to be observed (i.e. even the row/column re-ordering and the thresholding do not allow to infer clear driven/driving roles). It is worth remarking again that this figure corresponds to a thresholded matrix, i.e. the version that should ease visualization if any patterns existed.

fig. S16. Thresholded $T^{\dagger}$ matrices corresponding to different moments in the Twitter unfiltered data set. As expected, no country-to-country driving patterns emerge.

Additionally, to have a complete picture, we include two figures (see below) with the actual time series (volume). Each set of plots (left, right) has 4 panels: global volume, volume for US (top contributor), Kuwait (the country in the middle, in terms of volume), and Nigeria (last included, in terms of volume).

The set of plots on the left represent raw volume at a resolution of 600 seconds (less noisy), the one on the right at the dominating time scale (45 sec).

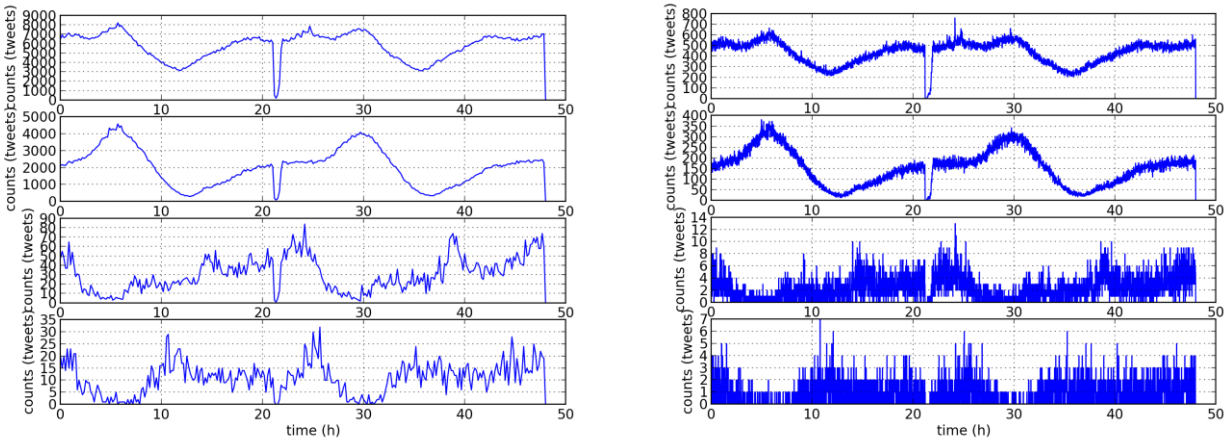

fig. S17. Raw time series for Twitter unfiltered stream for Δt = 600 s and Δt = 45 s (left and right, respectively). In both cases, global (top), US (second), Kuwait (third) and Nigeria (bottom) time series are represented.

Some interesting insights can be drawn from the raw time series:

- In general, we observe circadian rhythms for these countries (2nd to 4th panel); the global panel (top) also shows such rhythm --as it is dominated by US.

- There is a data gap at t ~21h, spanning approximately one hour. This is perhaps due to a storage problem in the Decahose system where the data were collected from. Because we always consider $\omega = 6$ hours, overall results are not affected.

- Because we are dealing with an unfiltered data set, the circadian rhythms are naturally shifted (different time zones). Note that Kuwait and Nigeria (3rd and 4th panels) seem to be advanced with respect with US (i.e. the trend in Kuwait at $t = 0$ is the same as US for $t = \sim 6$). It is important to highlight that this shift can imply that STE detects spurious driving (which would translate into hot colors in a plot similar to those in Figure 3 of the main text). However, this doesn't affect the timescales, which are low and relatively constant whatsoever. And, while an information flow evolution plot (similar to Figure 3 in the main text) might wrongly represent spurious driving, the "dominance matrices" $T^{\dagger}$ (fig. S16) are telling us precisely that no real driving is actually happening. That's precisely why the order parameter θ is absolutely relevant.

Finally, we would like to stress that the window width choice for this counterexample (ω = 6h) represents a harder test for our framework, because a smaller time window should detect shorter events (say, the rise and fall of some news within hours).

F. Validation of results (III): Synthetic time series generation

This work is focused on the capabilities of STE as a viable tool to analyze empirical data. However, it is under a synthetic controllable setting that the capabilities and findings of our approach can be made explicit –that is, the existence of transitions in the time scales (slow-to-fast) and directional couplings (asymmetric-to-distributed) in systems that gather around some sort of collective action.

Here, we offer two synthetic examples to better show the relationship between the optimal time scale (τ), the overall activity (or volume) and the dynamic interaction. These, as entangled as they may seem in our empirical examples, are here clearly separated.

Part 1) Minimalist example: Disentangling volume and time scales (Δt).

The question: can we imagine two constant-volume scenarios that maximize STE at different Δt? The answer is yes:

Take a time series Y = constant (e.g. all 0s, or all 1s, it won't make a difference)
Now imagine two time series X_i that have the same activity volume:

$$X_1 = \{1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\}$$
$$X_2 = \{1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\}$$

(each tick is a generic time unit, and we test $\Delta t = 1$ and $\Delta t = 2$, with m = 2)

Note that the volume is constant system-wide and even subsystem-wise (i.e. overall activity is constant, and is constant also if you check individual time series).

Results:

<u>System X_1,Y</u>

$$\Delta t = 1 \qquad T_{x1,y} = \begin{pmatrix} 0 & .014 \\ 0 & 0 \end{pmatrix}$$

$$\Delta t = 2 \qquad T_{x1,y} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

We infer from this that the optimal time-scale in terms of STE is $\tau = 1$

System X_2,Y

$$\Delta t = 1 \qquad T_{x2,y} = \begin{pmatrix} 0 & .007 \\ 0 & 0 \end{pmatrix}$$

$$\Delta t = 2 \qquad T_{x2,y} = \begin{pmatrix} 0 & .08 \\ 0 & 0 \end{pmatrix}$$

We infer from this that the optimal time-scale in terms of STE is $\tau = 2$

Therefore it is proved that *it is the dynamical interplay between the system's units that modifies the optimal time scale*. While the opposite is not always true: changes in volume do not *necessarily* mean changes in the dynamical interaction: just think of *Y* as an always-increasing time-series (instead of constant), and the result would be exactly the same (because *Y* would be translated into a single symbol, i.e. a constant symbolized series, rendering again the reported results).

Under this extremely simple example there is an implicit idea: one could come up with a synthetic model –which would allow for extensive testing in unlimited scenarios– if there were some efficient way to generate time series with some prescribed permutation entropy (which is beyond the scope of the work, but see ref. [22]).

Part 2) Nonlinear Lorentz oscillators: Time scales, volume, and dynamical coupling

Another way to approach the entanglement between time scales, volume, and dynamic interaction is to revisit the original Symbolic Transfer of Entropy paper by Staniek & Lehnertz [21]. In their paper, the authors propose a couple of synthetic benchmarks on which they demonstrate STE's virtues. It seems then natural to get inspiration from them here too.

Here, we monitor two Lorentz oscillators with varying levels of coupling, and see if changes in the time scale are observed. For our purposes, there are two stages we are interested in (but the possibilities are virtually infinite): (a) oscillators *A* and *B* are completely independent from each other (uncoupled, $\beta = 0$); (b) we increase the coupling strength such that oscillator *A* drives *B* (asymmetric coupling, $\beta > 0$. The results can be seen in fig. S18.

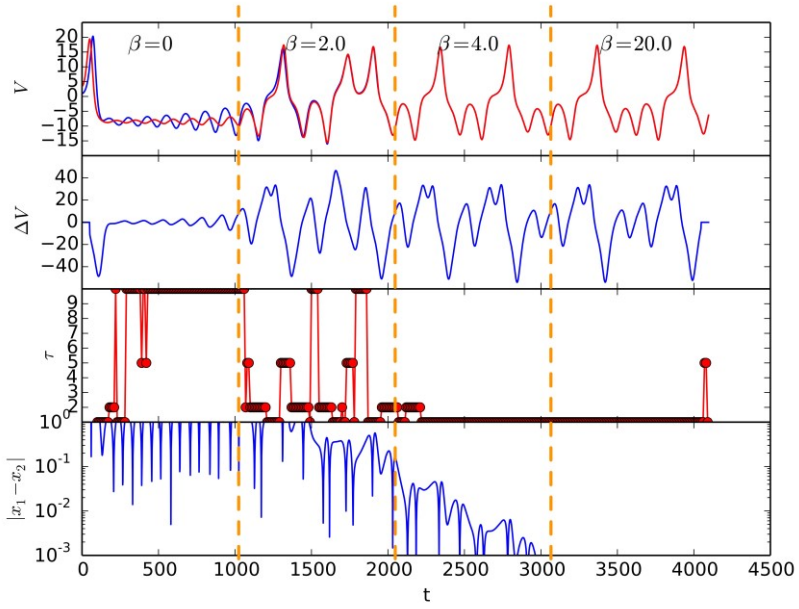

fig. S18. Evolution of two nonlinear systems under four changing scenarios: from dynamic independence (β = 0) to strong asymmetric coupling (β = 20.0).

This figure represents the evolution of the x-dimension of both oscillators for 4096 integration steps. As indicated by vertical lines, the unidirectional coupling strength β from oscillator *A* to *B* has been progressively increased, each 1024 steps. STE has been calculated for the whole system, with $w = 100$, $m = 5$ and different resolution levels, $\Delta t = \{1, 2, 5, 10\}$.

In the top panel we can see the evolution of both oscillators (blue: *A*; red: *B*). The second panel shows the variation in volume from window to window ($\Delta V_t = V_{w(t)} - V_{w(t-1)}$). The third panel shows the timescale at which STE is optimal (i.e., τ). Finally, the fourth plot shows the absolute difference between the values from oscillators A and B, as a (rough) proxy for their level of synchronization (note the log scale).

Some lessons from the plot can readily be extracted. First, the time scale is very sensitive to the coupling of the subsystems. When *A* and *B* are uncoupled, the optimal time scale is the slowest (τ = 10); the system immediately reacts to the introduction of slight coupling strength β, the time scale drops to 2 – then behaves erratically to some extent, as the system is not yet strongly coupled nor fully synchronized. Note that the sudden decrease in τ appears to happen slightly after the coupling strength is increased: this is an artifact of our representation, as we always report results for the *past* sliding window. That is, each red dot in the plot represents the corresponding τ for the *previous w* points of the time series.
For the stages with large β both systems become fully synchronized and the time scale remains low. *It is the dynamic relationship between both oscillators that determines* τ.

It is interesting to see the evolution of Δ*V* and τ (2nd and 3rd panels). The "variation in volume" Δ*V* is mostly constant during the first stage; after it, volume is rapidly changing (large Δ*V*). Out of visual inspection, it is possible to observe decreases in τ when volume is growing (Δ*V* > 0); and yet, it is

possible to observe also the opposite, i.e. τ *can decrease with volume decreases* ($\Delta V < 0$). The straightforward relationship (increased activity implies short timescales, and vice versa) is –now via dynamical models– refuted.

On a noisy τ: During the second stage of the dynamics (unidirectional weak coupling, $\beta = 2.0$) it is possible to observe some fluctuations of the optimal time scale τ. These can be explained from the chaotic trajectories of the units, which, under the mentioned weak regime, sometimes are close to synchronization (and sometimes not: note also the fluctuations in the bottom panel). STE (or any other measure, for that matter) is incapable of disentangling these different scenarios and reflects, at this very local level (keep in mind that $w = 100$ only). On the positive side, this brings some light to the accuracy and sensitivity of the measure; on the downside, it provides at times a noisy result –which nonetheless is, to our understanding, quite robust if observed for a longer time span. That is, if we observe the overall differences in τ for the periods $t < 1024$ and $1024 < t < 2048$, it becomes clear that the time scale has undergone a transition.

Having said that, we finally wish to highlight that these local fluctuations can be smoothened if a longer w is considered: see fig. S5. This comes at the cost of a less sensitive measure –the time scale decrease is observed at later stages.