What an entangled Web we weave:

An information-centric approach to socio-technical systems

Markus Luczak-Roesch, School of Information Management, Victoria University of Wellington, New Zealand

Ramine Tinati, Electronics and Computer Science, University of Southampton, United Kingdom

Kieron O'Hara, Electronics and Computer Science, University of Southampton, United Kingdom

Corresponding Author

Markus Luczak-Roesch, markus.luczak-roesch@vuw.ac.nz

Keywords

Socio-technical Systems, Information Theory, Networks and Communities, Temporal Data Mining, Collective Intelligence, Social Machines, Augmented Intelligence, Citizen Science, Online Communities, Information Dynamics

ABSTRACT

Motivated by the increasing amount of voices who ask for careful consideration of what context-rich data analysis methods can tell us about the activities of human collectives, we contribute an argumentation that employs a dialectic of literature on the philosophy of truth and science as well as analytical methods for the study of information diffusion, Web graphs and social networks in order to make a case for changing the current view to the actions of human collectives in the digital. We strengthen our meta argument by a case study about one particular method that breaks with the *causality assumption* that is inherent in many of today's methods and allows to capture novel dimensions of complexity of information sharing from a macroscopic cross-system perspective. We discuss whether this kind of analysis may generically suit to underpin the field of socio-technical systems with a novel information-centric theory.

INTRODUCTION

What is the structure of the World Wide Web? A question that had a relatively simple answer to it about 15 years ago (Broder et al., 2000), has no answer to it anymore today. Or at most the unsatisfying answer: We don't know! In this article we will unpack arguments why this limited view has an impact beyond the focused space of information retrieval and Web Science research, and has wide implications for the broader field of information systems research as well as computational social science.

Here we contribute an argumentation that employs a dialectic of literature on the philosophy of truth and science as well as analytical methods for the study of information diffusion, Web graphs and social networks in order to make a case for changing the current view to the actions of human collectives in the digital. To strengthen the point, we present a case study

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2789v1 | CC BY 4.0 Open Access | rec: 8 Feb 2017, publ: 8 Feb 2017

about one particular method that breaks with the *causality assumption* that is inherent in many of today's methods and allows to capture novel dimensions of complexity of information sharing from a macroscopic cross-system perspective. This case links back to the meta-argument in the article – the call for a more holistic approach to information in socio-technical systems – as it transcends the common boundary between the "token view" and the "information in the syntax view" in the information systems field (McKinney & Yoos, 2010; Lee, 2010).

WHEN SOCIALLY DETERMINED NETWORK MODELS FALL SHORT

In the 25 years since its inception, the World Wide Web has evolved from a hypermedia system with rather low information sharing dynamics to a space where a) content is shared at very high (and still growing) rate and b) links between Web content are increasingly implicitly emerging from common metadata (e.g. categories) or patterns in the actual content such as hashtags or the mentions of usernames. We see a particular tendency in many of today's analytical methods that are applied to socio-technical systems of the kind found on the World Wide Web or in modern organizational information systems in order to capture these increased dynamics: they assume that there must be some retrievable snapshot structure underlying the actions of the human collective that can then be used to infer causality (call this the *causality assumption*).

The analytical toolbox typically relies on structures of explicit relationships between entities along which information can "diffuse" (e.g. blog sites interlinked by blogroll features or users forming a following or friendship graph) (Gruhl et al., 2004; Adar and Adamic, 2005; Leskovec et al., 2007). An actual diffusion process is then represented as a directed overlay network with each edge in the overlay network being directed from the "infector" node to the "infectee" node. Evidence for an infection is inferred from contextual features of the

underlying network (Leskovec, 2007 & 2009; Goel, Watts and Goldstein, 2012; Qu et al. 2014). With such models, the path of information through a system-specific context network can be traced quite accurately. However, while the causality assumption underlying these analytical methods may hold for selected systems, it is unlikely to hold for the entire macroscopic space of all systems that together form the World Wide Web.

The problem by example

To demonstrate this, we consider the example of digital disaster response. The earthquakes in Haiti and Nepal, the political crises in Congo and Somali, or the recent Ebola outbreak in West Africa are representative cases where dedicated Web applications combined with general social media platforms were used opportunistically to gather crisis-related information in order to respond more effectively. The information relevant for a particular crisis goes well beyond that shared by individuals to support crisis management directly. Depending on the platform on which the information is shared, it can be intended (e.g. contributions via an instance of the Ushahidi tool suite) or accidental that information is relevant to relief coordination (e.g. a micropost on Twitter about one or more cancelled flights from or to a particular airport could be relevant information for aid workers who have to reach a crisis region from outside of that area). This emphasizes that a) crisis related information does not necessarily reside on a single platform and b) even when selecting a particular platform to capture crisis related information, the relevance depends on the content rather than the social networks that that platform is supporting. As depicted in Figure 1, the collective action is more visible from the point of view of the dynamics through time of the content rather than by explicit social structure (Lee & Paine, 2015). We must expect the relevant information about an event that affects many individuals – particularly heterogeneous individuals who have only weak links with the majority of others affected – to overflow from any individual communication channel.

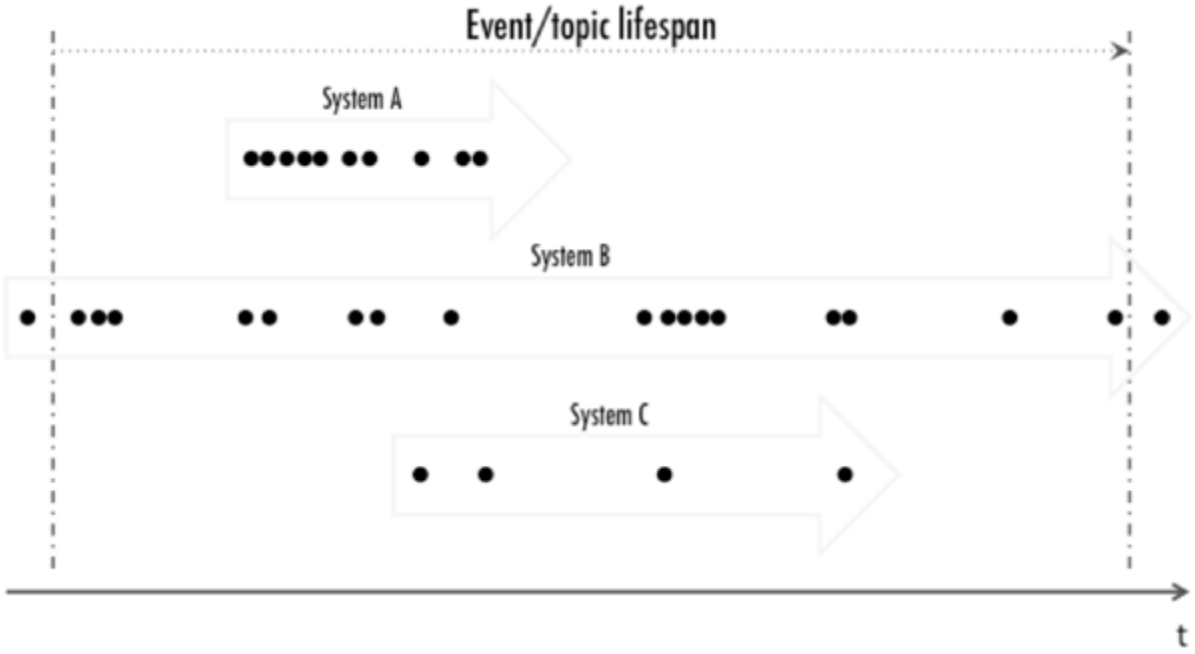

Figure 1: Information naturally resides in an ecosystem and is emitted at varying frequencies. The accumulated information that is relevant for an event or topic forms the implicit collective action related to it.

The intuition behind our argument is a simple one. For a human conversation, or collection of conversations connected by a particular topic, the media of those conversations are of secondary importance to the participants (though of course, recalling McLuhan's famous insight, this is not to say that the medium is irrelevant to understanding the message). Human interlocutors are swayed by many factors in their choice of medium for a particular communicative act, from opportunism, to habit, to the need for security and/or anonymity, to the devices in their (and their interlocutors') possession, to requirements for synchronous discussion, to the need to reflect and to gather information before communicating, to whether the interlocutors are communicating in official or private capacities, etc. Someone who needs to communicate with someone else will use whatever is to hand, rather than intentionally restrict the conversation to a single channel such as Twitter or email. A conversation made up of many communicative acts may therefore take place across many different media, and this

will be compounded when we aggregate conversations to try to set out an inclusive information picture. Many communicative acts will take place in face-to-face speech or in other unrecorded and unrecordable ways, and so these are not going to be captured. However, a method that can encompass a set of media, wider than an individual channel such as a social networking site or a microblogging site, will be better-placed to be inclusive in this way.

Furthermore, a group of conversations will be connectable via exogenous events. Again the intuition is simple – the occurrence of a major event, such as a crisis (an earthquake, say) will trigger a number of independent conversations which will use similar vocabulary and identifiers on similar timescales. The aggregate of these conversations may be of great importance to crisis managers or rescue workers, and once more to focus on individual sources of data with particular models of discussion embedded into particular information infrastructures misses a trick. The wider collective discussion is of interest – a series of conversations connected only by the basic relationships of *being about the same thing* and *taking place during a key time period*.

To be sure, it is easier, if we simply take the data from a single channel, to make contextual assumptions, for example about the connection between two communicative acts (it may be built into the infrastructure when C' is a reply to C, for example). And it is easier to manage data if we take it from a single channel. If we move across channels, then it may be that two communicative acts that share a vocabulary and are closely connected in time are merely coincidental. However, if we can construct maximally inclusive structures, even if these consist of both genuine conversations and coincidentally-related communications, then further analysis can bring contextual factors back into play to pare the structure down to

something more intentional. Traditional methods in this space tend to assume that the available data – for example from a social networking site – captures the complete conversation and its context. This is a handy assumption for researchers, but is unlikely to reflect the understanding of the participants.

Truth is not viral

This problematic situation has recently been described as a general "malformation" (Marcinkowski, 2015) of socio-technical systems. Critical voices emphasize that quantitative analytical methods relying on system-specific digital structures are in a general epistemological trap of misleading local maxima and a limited view of phenomena that retrospectively stand out as assumingly successful social action in the digital (Lerman, Yan and Wu, 2016; Cebrian, Rahwan & Pentland, 2016). And this trend of *viral truth* is theoretically challenged because any informational pattern that is passed on through a socio-technical system can at most be regarded as *suppositio materialis* (Tarski, 1944). The relationships manifested in a diffusion network do not necessarily have any defining or designating reference to the knowledge object itself and the relevant dynamics bypass traditional notions of verification and falsification (Popper, 1972). The resulting information flow has been called the 'post-truth' world.

Acknowledging complexity beyond pure reason: An Entangled Web?

Our aim, therefore, is to consider what it would look like to understand evolving information in abstraction both from the social networks through which they flow, and from the system-specific digital traces of the social context. As outlined before, such consideration of socio-technical systems implies that system-specific data is incomplete to describe the macroscopic state of the space of all relevant information. Unconventional or hidden relationships between information – that would usually appear as noise relative to the explicit social relationships

between the originators of the information – may be very influential despite, or independent of, the social structures created and curated by networking systems. Hence, we suggest exploring the possibility of separating the social context from the technological substrate to understand the Web's contribution qua abstract information space to the evolution of information. Whereas research on collective intelligence and human computation typically focuses on groups working explicitly or implicitly together towards a particular outcome and the coordination to optimize this (Malone, Laubacher & Dellarocas, 2009; Woolley et al., 2010; Quinn & Bederson, 2011), in this research the goal is to expose the resources available from accumulated activities of human users on the Web while minimizing the presuppositions about the communities or systems in which they take part.

A generic model called Transcendental Information Cascades has been proposed (Luczak-Roesch et al., 2015a; Luczak-Roesch et al., 2015c), conceived as networks of information co-occurrence dependent solely on time and inherent properties of pairs of content resources. This can be referred to as a transcendental method in Kant's sense of attempting to understand the conditions of knowledge itself (Kant, 1934). This means that not all such networks represent underlying purposeful activity – co-occurrence may simply be coincidence; but they do present a distinct set of properties of the macroscopic informational state of the Web or any Web-based information infrastructure. The relationships within and between such networks will reveal linking structures hidden from the system-specific point of view as well as temporal dependencies, describing **a kind of entanglement between Web resources**.

STUDYING ENTANGLEMENT PHENOMENA IN WEB-BASED INFORMATION SYSTEMS

We will now turn to the technicalities of Transcendental Information Cascades in order to demonstrate that there is already analytical capacity to capture the organic informational state of socially constructed information in systems of systems. This shall demonstrate what tools or techniques may help understanding of this **Entangled Web** independently of curated networks, and provide a reflection on two cases in which Transcendental Information Cascades are key to uncover otherwise hidden relationships between Web resources.

Formally a Transcendental Information Cascade is defined as a directed network, which is constructed by applying an arbitrarily-chosen but, once chosen, fixed set of information extraction algorithms to a chronological sequence of discrete content elements (see Figure 2 for an overview of the cascade construction and analysis process). Those content elements, for which the information extraction algorithms match one or more informational pattern, become the nodes of a particular cascade network. The matched patterns are termed cascade identifiers and form the variables for constructing a cascade network out of a dedicated set of content elements (an alternative configuration of the information extraction algorithms may lead to different patterns to be matched). Under the assumption that the time difference between all nodes is always positive, an edge exists between any two nodes that share a unique subset of cascade identifiers such that none of the identifiers is matched for any node published at a time between the two nodes to be linked. These networks of information co-occurrence are context-free in the sense that no global feature set or pre-existing structure is exploited for their generation, including any assumptions made as part of a social networking architecture. Edges only result from the comparison of pairs of resources. That does not mean (a) that no context exists, (b) that it is unimportant, or (c) that it should not be taken into

account in the investigation of the cascade, only that we need to construct the cascade as an antecedent step, because the structures we investigate will be biased if we smuggle assumptions about their context into their construction. Rich context can be added after the cascade construction to weight edges or label nodes for example.

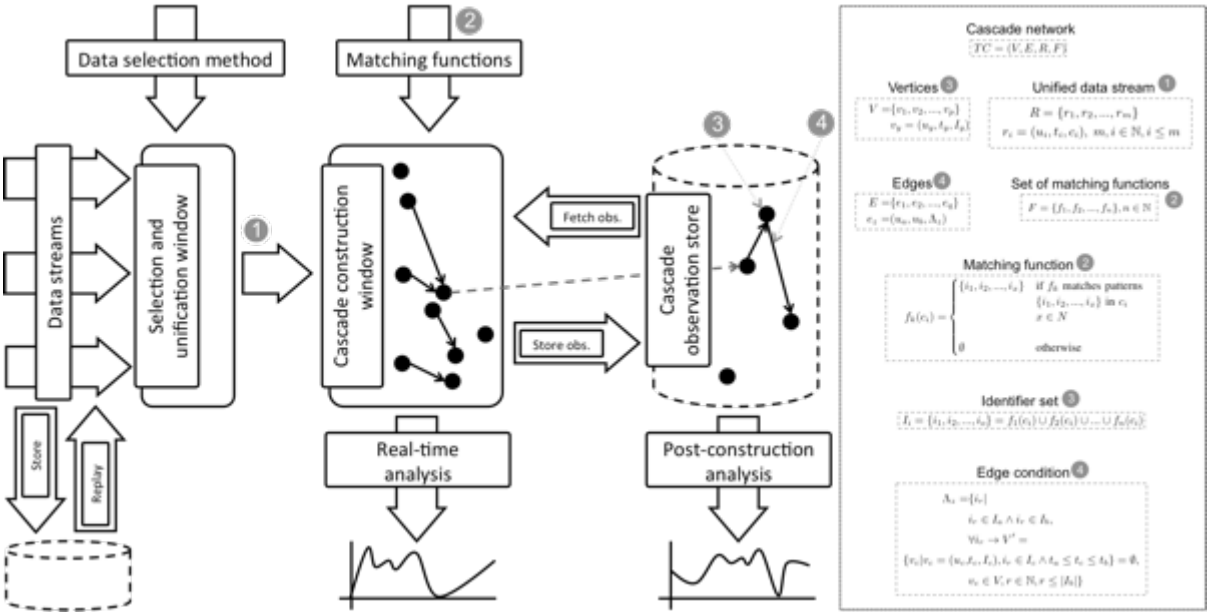

Figure 2: The construction of Transcendental Information Cascades is configured by the input data sources providing time stamped content sequences, a data selection method to to filter out a unified content feature from the data (e.g. free text content), and one or more information extraction algorithms that are applied as identifier matching functions to this content. Analysis can be done in real-time on the resources currently observed or on the stored cascade graphs. Replaying historic data allows for experimenting with new selection methods and matching functions or to study effects that would have been observable when a particular simulated situation would have occurred (e.g. by injecting synthetic data into the input stream).

This cascade model yields different structures depending on both the data at hand and the information extraction algorithms applied, which serve to generate the particular cascade identifiers. Algorithms can be selected opportunistically, depending on (a) what is possibly significant, and (b) what structures are unlikely to be uncovered by more conventional methods; one could imagine searching for cascade identifiers within hashtags, URIs, quotes, topics, keywords, images, or even semantics and sentiments. Where the traces of co-occurring information appear to collide serendipitously, we can focus our further investigation. Even though edges are only created between directly consecutive content elements that share an identifying pattern, implicitly any resource is in interaction -- or entangled -- with any other resource in the cascade it belongs to.

Distinct properties of Transcendental Information Cascades

Central to our current understanding of the meaningful application of Transcendental Information Cascades are the following three scenarios: the use of a single cascade to study a process; the use of multiple cascades to understand the significance of different types of information; and a framework to tie multiple cascades together into a coherent overall picture.

Sampling relevant time windows by assessing intra-cascade properties

The nature of Transcendental Information Cascades as directed networks preserving a concise set of informational patterns for each node allows well-established quantitative methods to be used to capture structural as well as informational properties of socially constructed information traces. The benefit of the approach is that fundamental low-level analytical methods can be used, so that the system-specific context inherent to the analysis (e.g. case specific feature sets) is reduced, allowing for unbiased discovery of significant patterns across (a cross-section of) the Web, rather than within (and therefore illicitly assuming the centrality of) particular restricted well-behaved and well-understood milieu.

Our analyses seek to determine where significant bursts of structural and/or informational patterns kick-off or fade away, indicating the emergence or shift of an underlying exogenous phenomenon and providing a trigger for sampling a particular subset from the overall content element sequence for more detailed inspection involving more context.

Mining implicit coordination by assessing inter-cascade properties

Based on these intra-cascade properties can we expand the view to inter-cascade properties? In particular, given that different configurations of information extraction algorithms applied to the same sequence of content elements will result in alternative cascade networks, the big question to be addressed is: Which of the resulting networks is the most appropriate representation of some underlying exogenous event or activity? Or phrased differently: What is the appropriate information extraction algorithm orchestration to capture an implicit collective action?

An answer to this question would provide the basis for devising an adaptive approach to cascade construction. Entropy measures reflecting the distribution of cascade identifiers can be used to determine which matched informational patterns are associated with a certain degree of randomness. This then allows refinement of the information extraction algorithms by excluding certain patterns from consideration (e.g. specific words or hashtags used for spam on social media which tend to tie together information randomly and not to reflect populations' priorities). Furthermore, bursty periods of different cascade networks can overlap (or be completely disjoint) indicating a relationship between (or independence of) the extracted cascade identifiers. Where there is a relationship, we can then concatenate selected information extraction algorithms to derive another alternative cascade network.

Detecting bursts along multiple axes by locating information in a multi-dimensional space

Detecting bursts of activity is a suitable means to infer exogenous events underlying socio-technical systems but it is typically focused on individual information streams (Kleinberg, 2003; Barabasi, 2005). If we model cascades of information co-occurrence to describe the global interconnected informational state in a socio-technical system, we can represent information in a multi-dimensional space so that we can see bursts occurring along different axes. Preserving the context-free nature of the approach, three dimensions are the natural base for this representation: (1) time; (2) an index of all unique cascade identifier sets extracted from data (reflecting the chronological order in which identifier sets are found); (3) an index for each unique identifier set which is incremented with each occurrence of the respective set over time. Adding context allows us to scale the number of dimensions variably (e.g. adding further dimensions for the system in which particular information occurred or the human individual who shared it). It may be, for example, that we would want to include a geographical dimension, because we are interested in the specific viewpoints of the heterogeneous set of actors able to influence a situation (Cebrian et al 2016); recall the events of the so-called "Twitter revolution" in Iran in 2009, when – thanks to over-reliance on the use of data from a single channel, Twitter – the prospects of the revolution's success were dramatically over-estimated as most relevant Twitter traffic turned out to be supportive tweets from the US and the UK, and, as (Honari 2015) put it, "various areas of interest to Iranian users have been neglected or ignored" in the literature.

This projection into a three-dimensional space allows the identification of (a) periods when new unique identifier sets are created at high frequency and (b) periods when particular identifier sets burst. While this space seems to naturally diverge over time from a macroscopic viewpoint, adding the cascade links to the visualization reveals ties across

individual information streams allowing the tracing of time-persistent dependencies that would be otherwise hidden.

Applications of Transcendental Information Cascades

The motivating use case described earlier was the study of digital disaster response as a collective phenomenon, but there are many cases where a macroscopic view of the accumulated information sharing has more value than the architecturally amplified activities of individuals on the Web. Following guidelines for case based research (Benbasat, Goldstein & Mead, 1987; Eisenhardt, 1989), we investigated the application of Transcendental Information Cascades to understanding real-world cases dependent on cross-channel communication: online citizen science and editing activities in Wikipedia. These cases suggest that Transcendental Information Cascades may provide a unique way of underpinning the field of socio-technical systems with a distinct information-centric theory.

Citizen science: Coordination by content

Online citizen science is a blueprint of the trending hybrid computing approach, coupling state-of-the-art artificial intelligence with human computation, to enable interested people to tackle problems in scientific research that are impossible to solve in a purely computational fashion. The Zooniverse, for example, is the world's largest multi-project citizen science platform, with over 1.3 million volunteers contributing to projects from various domains such as astrophysics, biology or digital humanities amongst others. The platform gained popularity as the source of numerous citizen-led discoveries made after participants had branched out beyond the immediate system-generated constraints, discussing outliers and making other remarkable serendipitous observations while performing the crowdsourcing task. Information sharing on those platforms often evolves to become domain-specific and goal-oriented. Hence, supporting this domain-specific information sharing around the objects examined as part of the crowdsourcing task has become part of the core of many citizen science systems.

However, from the point of view of research methods in information dynamics, these systems are very peculiar with respect to the online communities they form. They typically do not feature explicit social networks and the community structures that emerge implicitly are highly fluid and dependent on many aspects of context (Luczak-Roesch et al., 2014).

Transcendental Information Cascades were applied to a dataset representing content sharing on the Planet Hunters project hosted on the Zooniverse (Luczak-Roesch et al., 2015a; Luczak-Roesch et al., 2015c). Four different information extraction algorithms based on string matching using regular expressions were tested on this dataset in order to derive alternative cascade networks: (1) hashtags; (2) matching of content that refers to specific object identifiers related to the images shown in Planet Hunters; (3) matching of identifiers used by the Planet Hunters community to refer to objects in external astrophysics databases; (4) URIs. The studies of the resulting cascade networks revealed that only the information extraction algorithms 2 and 3 were suited to be combined without further adaptation. The cascades derived by applying these methods naturally showed patterns of disjoint local phenomena, which were correlated in time. Meanwhile, cascades based on hashtags tended to be either single identifier cascades or consist of multiple roots that merged and diverged to form a single massive connected component from which little useful information could be extracted. URI-based cascade networks tended to feature a significant fraction of independent cascades in which one particular identifier set recurred repeatedly. Hence hashtag and URI cascades would need to be refined first, until the intra-cascade properties indicated the same distinctiveness as the other two approaches. Note that, in this case at least, the identifiers that were already built into the system were less insightful compared to expressions that evolved within the community (e.g. KID identifiers) and consistent with our argument to move beyond system-specific features to uncover interesting relationships.

Wikipedia edits: A source of temporal relationships

Wikipedia represents a network of human-curated, moderated, and maintained articles, which over time have become the largest encyclopedia in existence. The variety of social processes in Wikipedia – from managing vandalism, to ensuring quality and consistency in the knowledge base, and even to detecting gender imbalances – allows us to consider it as more than just a network of explicitly linked articles. Implicit structures emerge from coordinated or sometimes just accumulated activities of volunteers, but can become explicit if the community approves them to be useful as exemplified by WikiProjects, an effort to form sub-communities in order to increase the quality of domain-specific article sets. For Wikipedia, a core challenge is to discover and in certain situations support such emergent phenomena effectively within the vast amount of user and machine-generated data. Every second, hundreds of articles are created or revised, edits are overwritten or reverted, abuses are reported, and discussions take place. This stream of activities represents the digital traces of collective human action, and to that end, studying these streams reveals temporal relationships between articles that would remain hidden otherwise and promises to provide insight into the underlying social activities of such a system from a novel angle.

As an example of the potential for progression from context-free cascade construction to context enrichment for interpretation and sense making, let us consider the evolution of Wikipedia edit logs, applying a string matching function to the text associated with each Wikipedia revision entry. The matching function uses a regular expression to identify trigram noun phrases to match entities like "The White House", "Barack Hussein Obama II" or "The Empire State Building" for example. In this situation Transcendental Information Cascades form a network of articles, linked together by the shared identifier found within the edit revision text. By enriching the article edits with contextual knowledge about article categories from DBpedia (http://dbpedia.org) it was possible to find that this cascade network

represents meaningful article relationships not available within the explicit network of linked Wikipedia articles.

An analysis of the informational and structural properties as well as the general burstiness characteristics of the constructed cascades showed that they reflect both external events and phenomena inherent to the system (Tinati, Luczak-Roesch & Hall, 2016a; Tinati, Luczak-Roesch and Hall, 2016b). In particular, a burst of activity was observed featuring a series of edits made within a short duration of time beginning with identifiers found in edits on the article about Edward Snowden. The cascade then branched out to span across many other articles incorporating various identifiers related to Edward Snowden's life. A detailed inspection of the time frame when the cascade emerged showed that it coincided with a presentation given by him at the SXSW conference. In other words, a relationship between an external phenomenon and a short, bursty cascade of edits within Wikipedia, which would not have been available to a more contextualized investigation, was uncovered using the method. In similar vein, we were also able to observe more local phenomena, such as a pathway found around the identifier: "U.S. District Court". This cascade extended over a longer period of time, linking articles and identifiers related to same-sex marriage in the United States, which led to an editing debate within the Wikipedia community around articles featuring lists of U.S. state laws on same-sex unions. Here, in contrast to the cascade from Snowden's talk, we were able to observe the frequent re-occurrence of articles within a single pathway indicating back-and-forth editing activity – an edit war – between Wikipedia editors.

Synthesizing the case of Transcendental Information Cascades

Both applications show how the construction of Transcendental Information Cascades reveals implicitly collective action within a stream of activity based on information co-occurrence independently from assumptions about prior structure or connectivity. This suggests that the

method has application for the detection of influence of exogenous phenomena as well as temporal contagion within socio-technical systems, underlining that these contain social groupings, susceptible to influence from the full range of social contexts and social networks in which individuals take part, not simply the specific medium, platform or architecture from which data can be harvested.

To reiterate, none of this is meant to imply that data about, or gathered from, social networks is unimportant – far from it. But some extra input is required to understand what sort of intelligence is detectable within a socio-technical system as a whole independently of assumptions about social mechanisms for its delivery (and of course this independence is earned at the cost of restricting our use of these assumptions about mechanism, at least as we construct the global information space). This view responds well to the argument that often social networks and social network analysis are not used in a scientifically rigorous sense in the information systems field (Lee, 2010).

Transcendental Information Cascades may be a complement to analyses that exploit rich contextual features as well as more complex a priori modelling or clustering of information (Shahaf et al., 2013; Shahaf et al., 2015). As the examples have shown, sometimes the necessary contextual data is not available (or, under a different privacy regime, may not be accessible), in which case alternative techniques such as the ones proposed here would anyway be required and welcome. The only general relationship presupposed is temporal precedence and the key subjects of interest are bursts (Kleinberg, 2003; Barabasi, 2005; Kumar et al., 2005), the low-level model prevalent in almost any temporal data mining approach (Mei and Zhai, 2005; Subašič & Berendt, 2013). The transcendental understanding of cascades, following our Kantian theme, is skeptical about apparent causal roots, rejecting

406 the ready-made etiology contained in social network data and focusing instead on the
407 narrower supporting base of whatever is detectable from time order and syntactic/semantic
408 coincidence. This attempt to devise an information-centric theory for socio-technical systems
409 enables a macroscopic view to the emergent output of complex social action by studying the
410 change of almost physical properties, which links it to social entropy theory (Bailey, 1990;
411 Bailey, 2006). This differentiates Transcendental Information Cascades from the system-
412 centric perspectives commonly referred to in Social Computing and Computer-supported
413 Cooperative Work (Grundin, 1994; Parameswaran & Whinston, 2007).
414
415 A Transcendental Information Cascade can be regarded a model that spans two of the four
416 views of information presented by McKinney and Yoos (2010). The low-level matching of
417 patterns for cascade construction basically means looking for small but meaningful units in
418 data sequences and reflects the "token view" (McKinney & Yoos, 2010; Lee, 2010). The
419 mechanism to add relationships between "temporally coincident" (Jung, 1952) occurrences of
420 those tokens lets the model transcend to the higher-order "information in the syntax view"
421 (McKinney & Yoos, 2010; Lee, 2010). It is this step that allows one to say that a
422 Transcendental Information Cascade channels and preserves information across time, which
423 has the potential to be the unique feature of the approach. It means that a Transcendental
424 Information Cascade has storage and transfer capacity, and as a result is an important concept
425 particularly for distributed communities which may have few communally-created
426 information storage facilities capable of allowing access to information in a timely manner at
427 the point at which it is needed. Some, but not all, input signals become output signals, so a
428 body of information can evolve over time; information loss may correspond to information
429 ceasing to be current, or alternatively a cascade might branch to create divergent cascades
430 whose combined capacity may make up for apparent local losses. All this requires

measurement and understanding, but many of the tools are readily to hand. When cascade paths collide at a particular point in time, the tools of information theory (Shannon, 1949; Kullback, 1968) can be used to understand the properties of the collision and the nature of the resulting entanglement as entropy will increase or decrease for example. Motifs of the network structure or the entropy over time can be aggregated into states, which have their own theoretical-analytical apparatus such as Markov models and finite state machines (Anick et al., 1982; Rabiner, 1989).

CONCLUSION

The aim of this work was to provide insight into a number of factors. First of all, there is the way in which the Web facilitates information evolution, abstracted away from the federation of co-created socio-technical systems (and walled gardens) whose aggregation we are accustomed to call, loosely, "the Web". We argue that it is important to minimize the number of assumptions we make about the social context of information evolution – not because we do not believe that social context plays a highly significant role, but rather to derive important social relationships from the information evolution of the Entangled Web, without reproducing the assumption that existing data from networking and sharing sites exhausts the relevant context (Facebook gives you the complete picture). This view provides an alternative, less powerful, less context-dependent, and potentially less deluded perspective on Web-based information systems in general and may be up for debate in the field as a kind of "box-breaking research" (Alevsson & Sandberg, 2014), raising a whole set of new questions about how we model and study emergent socio-technical systems.

Secondly, we hope to be able to understand the Web as a wider phenomenon than siloed representations in specific networks tend to imply, a coherent phenomenon, a chord rather than its arpeggiated components, expressing its state at a time by quantifying the information

represented (and its dynamics). This is valuable for research on Social Machines (Berners-Lee, 2000; Hendler & Berners-Lee, 2010), as characterized by Smart et al. (2014) as "Web-based socio-technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realization of system-level processes." Our work contributes insight into the *organic* "system-level processes", so that the computation of Social Machines becomes the output of this kind of analysis (Luczak-Roesch et al., 2015b), rather than one of the inputs, and no assumptions are made about Social Machines as marooning themselves on particular channels (on which we happen to have the data). This suggests that there exists an interesting new generic phenomenon in socio-technical systems that we call **not necessarily coordinated collectives**. A not necessarily coordinated collective is a group of people treated as equal contributors to an accumulated activity stream, regardless of any pre-defined real or virtual relationships between those people or their contributed content.

The models and experiments we have discussed here are of course very small steps on what will be of necessity a long journey of research, experimentation and much more complex macroscopic and microscopic investigation. For instance, how do we determine the informational properties of any possible pairs of resources on the Web; find the best partitioning of a cascade network into the minimum number of non-nested sub-structures to derive an aggregated state machine representation; mine the collective intent of the people involved in the contents of particular Transcendental Information Cascades? We need the capacity to index and search for, not only documents and data (Brin & Page, 1998; Broder, 2002), but also Transcendental Information Cascades themselves, enabling us to understand how information dynamics facilitate and are facilitated by procedural knowledge. In the end,

such understanding will have engineering repercussions, as we seek to create the conditions for the effective creation of knowledge using Web technologies.

ACKNOWLEDGEMENTS

This work is partially supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/2 and comprises the Universities of Oxford, Southampton, and Edinburgh.

REFERENCES

Adar, E. and Adamic, L.A., 2005, September. Tracking information epidemics in blogspace. In Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence (pp. 207-214). IEEE Computer Society.

Alvesson, M. and Sandberg, J., 2014. Habitat and habitus: Boxed-in versus box-breaking research. Organization Studies, p.0170840614530916.

Anick, D., Mitra, D. and Sondhi, M.M., 1982. Stochastic theory of a data-handling system with multiple sources. The Bell System Technical Journal, 61(8), pp.1871-1894.

Benbasat, I., Goldstein, D.K. and Mead, M., 1987. The case research strategy in studies of information systems. MIS quarterly, pp.369-386.

Bailey, K.D., 1990. Social entropy theory. SUNY Press.

Bailey, K.D., 2006. Living systems theory and social entropy theory. Systems Research and Behavioral Science, 23(3), pp.291-300.

Barabasi, A.L., 2005. The origin of bursts and heavy tails in human dynamics. Nature, 435(7039), pp.207-211.

Berners-Lee, T., Fischetti, M. and Foreword By-Dertouzos, M.L., 2000. Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. HarperInformation.

Brin, S. and Page, L., 1998. Anatomy of a large-scale hypertextual web search engine. In Proceedings of the 7th International World Wide Web Conference (Brisbane, Australia, Apr. 14–18). pp. 107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., 2000. Graph structure in the web. Computer networks, 33(1), pp.309-320.

Broder, A., 2002, September. A taxonomy of web search. In ACM Sigir forum (Vol. 36, No. 2, pp. 3-10). ACM.

Cebrian, M., Rahwan, I. and Pentland, A.S., 2016. Beyond viral. Communications of the ACM, 59(4), pp.36-39.

Eisenhardt, K.M., 1989. Building theories from case study research. Academy of management review, 14(4), pp.532-550.

Goel, S., Watts, D.J. and Goldstein, D.G., 2012, June. The structure of online diffusion networks. In Proceedings of the 13th ACM conference on electronic commerce (pp. 623-638). ACM.

Grudin, J., 1994, May. Computer-Supported Cooperative Work: History and Focus. Computer 27, 5, 19-26.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A., 2004, May. Information diffusion through blogspace. In Proceedings of the 13th international conference on World Wide Web (pp. 491-501). ACM.

Hendler, J. and Berners-Lee, T., 2010. From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence, 174(2), pp.156-161.

Honari, A., 2015. Online social research in Iran: a need to offer a bigger picture. CyberOrient, 9(2), http://www.cyberorient.net/article.do?articleId=9687.

Jung, C.G., 1952. Synchronicity: An Acausal Connecting Principle. In Vol. 8. of the Collected Works of CG Jung. Princeton University Press, 2010.

Kant, I., Critique of pure reason. Translated by Norman Kemp Smith. London Macmillan 1934.

Kleinberg, J., 2003. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7(4), pp.373-397.

Kullback, S., 1968. Information theory and statistics. Courier Corporation.

Kumar, R., Novak, J., Raghavan, P. and Tomkins, A., 2005. On the bursty evolution of blogspace. World Wide Web, 8(2), pp.159-178.

Lee, A.S., 2010. Retrospect and prospect: information systems research in the last and next 25 years. *Journal of Information Technology*, *25*(4), pp.336-348.

Lee, C.P. and Paine, D., 2015, February. From the matrix to a model of coordinated action (MoCA): A conceptual framework of and for CSCW. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 179-194). ACM.

Lerman, K., Yan, X. and Wu, X.Z., 2016. The" majority illusion" in social networks. PloS one, 11(2), p.e0147617.

Leskovec, J., Backstrom, L. and Kleinberg, J., 2009, June. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 497-506). ACM.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N.S. and Hurst, M., 2007, April. Patterns of Cascading behavior in large blog graphs. In SDM (Vol. 7, pp. 551-556).

Luczak-Roesch, M., Tinati, R., Simperl, E., Van Kleek, M., Shadbolt, N. and Simpson, R.J., 2014, June. Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science. In ICWSM.

Luczak-Roesch, M., Tinati, R., Van Kleek, M. and Shadbolt, N., 2015, August. From coincidence to purposeful flow? properties of transcendental information cascades. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 633-638). IEEE.

Luczak-Roesch, M., Tinati, R., O'Hara, K., & Shadbolt, N. 2015, February. Socio-technical computation. In Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (pp. 139-142). ACM.

Luczak-Roesch, M., Tinati, R. and Shadbolt, N., 2015, May. When resources collide: Towards a theory of coincidence in information spaces. In Proceedings of the 24th International Conference on World Wide Web (pp. 1137-1142). ACM.

Malone, T.W., Laubacher, R. and Dellarocas, C., 2009. Harnessing crowds: Mapping the genome of collective intelligence.

Marcinkowski, M., 2015. Data, ideology, and the developing critical program of social informatics. Journal of the Association for Information Science and Technology.

McKinney Jr, E.H. and Yoos, C.J., 2010. Information about information: A taxonomy of views. *MIS quarterly*, pp.329-344.

Mei, Q. and Zhai, C., 2005, August. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 198-207). ACM.

Parameswaran, M. and Whinston, A.B., 2007. Research issues in social computing. Journal of the Association for Information Systems, 8(6), p.336.

Popper, K.R., 1972. Objective knowledge: An evolutionary approach.

Qu, Q., Liu, S., Jensen, C.S., Zhu, F. and Faloutsos, C., 2014, September. Interestingness-driven diffusion process summarization in dynamic networks. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 597-613). Springer Berlin Heidelberg.

Quinn, A.J. and Bederson, B.B., 2011, May. Human computation: a survey and taxonomy of a growing field. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1403-1412). ACM.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), pp.257-286.

Shahaf, D., Guestrin, C., Horvitz, E. and Leskovec, J., 2015. Information cartography. Communications of the ACM, 58(11), pp.62-73.

Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H. and Leskovec, J., 2013, August. Information cartography: creating zoomable, large-scale maps of information. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1097-1105). ACM.

Shannon, C.E., 1949. Communication theory of secrecy systems. Bell system technical journal, 28(4), pp.656-715.

Smart, P., Simperl, E. and Shadbolt, N., 2014. A taxonomic framework for social machines. In Social Collective Intelligence (pp. 51-85). Springer International Publishing.

Subašić, I. and Berendt, B., 2013. Story graphs: Tracking document set evolution using dynamic graphs. Intelligent Data Analysis, 17(1), pp.125-147.

Tarski, A., 1944. The semantic conception of truth: and the foundations of semantics. Philosophy and phenomenological research, 4(3), pp.341-376.

Tinati, R., Luczak-Roesch, M. and Hall, W., 2016, April. Finding Structure in Wikipedia Edit Activity: An Information Cascade Approach. In Proceedings of the 25th International

Conference Companion on World Wide Web (pp. 1007-1012). International World Wide Web Conferences Steering Committee.

Tinati, R., Luczak-Roesch, M., Hall, W. and Shadbolt, N., 2016, April. More than an Edit: Using Transcendental Information Cascades to Capture Hidden Structure in Wikipedia. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 115-116). International World Wide Web Conferences Steering Committee.

Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N. and Malone, T.W., 2010. Evidence for a collective intelligence factor in the performance of human groups. science, 330(6004), pp.686-688.