From Coincidence to Purposeful Flow? Properties of Transcendental Information Cascades

Markus Luczak-Roesch, Ramine Tinati, Max Van Kleek, Nigel Shadbolt
Electronics and Computer Science, University of Southampton, United Kingdom
Email: {mlr1m12||$rt506$||$emax$||nrs}@$ecs.soton.ac.uk$

***Abstract*—In this paper, we investigate a method for constructing cascades of information co-occurrence, which is suitable to trace emergent structures in information in scenarios where rich contextual features are unavailable. Our method relies only on the temporal order of content-sharing activities, and intrinsic properties of the shared content itself. We apply this method to analyse information dissemination patterns across the active online citizen science project Planet Hunters, a part of the Zooniverse platform. Our results lend insight into both structural and informational properties of different types of identifiers that can be used and combined to construct cascades. In particular, significant differences are found in the structural properties of information cascades when hashtags as used as cascade identifiers, compared with other content features. We also explain apparent local information losses in cascades in terms of information obsolescence and cascade divergence; e.g., when a cascade branches into multiple, divergent cascades with combined capacity equal to the original.**

Keywords*—*Information cascades, information theory, network analysis, content analysis.

I. Introduction

Since its advent, the Web has continually re-defined how content is shared. While sharing used to be person-to-person, the over 3 billion digitally-connected individuals now most routinely share content using social networks and sharing platforms that have made mass-sharing the norm. When content is now shared, it may be picked up serendipitously by one of thousands of people who may not know the original sender, and who may, in turn, re-share, re-purpose and re-target the content to a another social group, thus causing the process to repeat itself again. The resulting transmission of information from one person to the next, referred to as *information cascades*, *information flow* or *information diffusion* in interaction networks, has been studied in order to understand the dynamics of information sharing across various communities and in many different settings.

Initially, the evolving blogosphere was the setting for much of the work on information cascades on the Web, but more recently attention has shifted to online communities. Most Web-based online communities today support the construction of "virtual social networks", through friendship and follower-following links, as well as several standard network-based sharing and content promotion mechanisms, including explicit targeted resharing,"retweeting", "liking" and "favoriting" functions. The propagation of content along social network lines can be seen as the addition of explicit links between the shared resources. Existing research has focused on exploiting such link structures to determine whether a piece of information has been shared in the same context as other sharing activities (e.g. in direct response or motivated by), which we will refer to as conditional information sharing. Existing work devised accurate methods for representing such conditional cascades and even predicting their growth and evolution, through a comprehensive set of 36 features that are, to a large extent, dependent on system-specific functionality (e.g. availability of knowledge about the time spend looking at content) and some existing sub-network (e.g. social network) [1].

Prior studies of information cascades have focused on the modeling, interpretation, and prediction of conditional information sharing activities, based on features derived from social networks, personal data, and tracked user interactions with a system. There are many settings online, however, where such features are unavailable, either because they not public or just not supported by the platform. A second difficulty occurs when cascades cross system/platform boundaries, where few of such features are unable to span. For example, while the information gathering and mobilisation efforts of volunteers during crises such as the Haiti 2010 earthquake often centres about a primary artefact on a platform (such as an Ushahidi[1] map) other coordination activities that support it crossed Twitter, YouTube, social networks, and many other channels. Individuals did not only talk with each other along explicit social ties, but to others with no prior connection across broadcast channels (especially in time-critical situations when time to make decisions is rare). Consequently, the existing cascade approaches will always be uncertain about some of the relationships that may exist between shared information, especially when borders of systems are crossed. Relationships might be missed out because the trigger event for their creation is not represented within the microcosm under investigation (e.g. a word-of-mouth recommendation about a trending hashtag on a microblogging system causing people to use it).

We also ask the question if there is value – or ultimately conditionality – in relationships that might appear to be random, yet result from information being shared coincidentally (e.g. two people having the idea to invent the same hashtag in separate microblogging systems at exactly the same time)? System-specific and feature-rich methods must be incomplete to describe this macroscopic informational evolution and state of the Web.

This paper is contextualised with these questions and investigates properties of *transcendental information cascades*, which are inspired by Kleinberg's work on bursty characteristics of document streams [2], [3]. This method requires only

[1] www.ushahidi.com

the temporal order of content-sharing activities that happen on the Web, combined with inherent properties of the shared resources. Hence, it is configured by a stream of time-stamped resources to be analysed, and a set of matching methods applied to generate relational links from their inherent properties. We applied this approach to the active online citizen science project Planet Hunters, hosted on the Zooniverse platform. The system is an excellent example to study this organic information evolution: it does not support any typical social networking functionality; there is evidence that also no virtual social network emerged from people's interactions implicitly; content sharing contributes to cooperative hypothesis testing.

Using this method cascade generation, we conducted experiments to investigate the following research questions: **(R1) Which changes in the resulting cascades can be observed when parameters of the matching methods are varied, while the input data stays stable? (R2) Is there a characteristic structural difference in cascades generated by our transcendental model compared to other approaches for the construction of information cascades? (R3) What is the role of cascade motifs?**

Our results give insight into how the structure of transcendental cascades can be affected by changing parameters of the matching methods. We find significant differences in the structures and properties of resulting cascades when hashtags are used as cascade identifiers, compared with other content features. We also explain apparent local information losses in cascades in terms of information obsolescence and cascade divergence; e.g., when a cascade branches into multiple, divergent cascades with combined capacity equal to the original. This complements Kleinberg's work by adding insight into the adaptation of his approach for Web content and substantially informs about how to configure experiments on the structural and informational properties of such coincidental information sharing activities.

The remainder of this paper is structured as follows: first, we start with a brief overview of related literature. We then describe our model of transcendental information cascades and a generic method of constructing them. We then describe our sample dataset and analysis methods applied to test our cascade model, followed by experimental results. Finally, these results are discussed alongside a set of final questions we lay out for future work.

II. Related Work

According to Kleinberg et al. [2], bursts refer to periods of significantly high activity in continuous, time-stamped sequences of documents. They have become an accepted indicator for the appearance of a topic [4] and can be used to infer meaning by analysing the content in documents that belong to a particular burst. This method has significantly influenced research studying the temporal properties of human-generated digital content (e.g. [5], [6]) but also been related to studies of human behaviour at scale [7]. We expand upon this approach, that is, centred on bursts in flat document streams, and seek to understand the role of reoccurring substructures of branching and merging flows of information.

A. Information cascades

Information cascades have been used to model a variety of information sharing practices online, spanning, for example, information propagation across blogs, the viral spread of news, memes and other content online, and influence and reach in political campaigns, to name a few [8], [9], [10]. Cascades are typically modelled as dynamic networks [11]. One or more undirected sub-networks represent structures of explicit relationships between entities along which information could possibly diffuse (e.g. users forming a following or friendship graph) and the actual diffusion processes are represented as overlays over those networks [12], [13], [14], [1]. We, in contrast, assume that there is a natural information transmission capability on the World Wide Web that is not necessarily conditioned by any preexisting sub-network. This allows for abstracting the social context away from the technological substrate to understand the Webs intrinsic information cascades, considering not only local understanding of its use but also a global view.

B. Collective Intelligence

By contextualising our alternative model with purposeful collaborative work, we also touch the field of collective intelligence, human computation, and social computing. Work in this area typically focuses on the intelligence and problem solving capability that results from virtually organised groups working together towards a particular outcome, as well as on coordination methods for making such collective processes run effectively efficiently [15], [16], [17], [18]. We instead want to expose the intelligence that lies in information on the Web that is linked because of coincidence rather than pre-configured conditionality, or necessary *a priori* planning.

III. Coincidence in Information Sharing

Transcendental information cascades were introduced in [3] as a 4-tuple $TC = (V, E, R, F)$, representing a directed network consisting of a set of nodes, V, and edges, E. This network is constructed by applying a set of matching functions, $F = \{f_1, f_2, ..., f_n\}$ to a set of resources, $R = \{r_1, r_2, ..., r_m\}$, where every resource is given as $r_i = (u_i, t_i, c_i), i \in [1, m]$, where u_i is a unique identifier of resource r_i, shared at time t_i, with content c_i. Nodes in the network are those resources from R that contain a set I_i of one or multiple *cascade identifiers*. A cascade identifier is any unique informational pattern that is recognised by applying a *matching function*, $f_k \in F$, to either the content of a resource, or any other inherent properties of it. Nodes, V, are then given as $V = \{v_1, v_2, ..., v_p\}, v_y = (u_y, t_y, I_y)$, while edges are given as $E = \{e_1, e_2, ..., e_q\}, e_z = (u_a, u_b, \Lambda_z)$, with $I_i = \{i_1, i_2, ..., i_o\} = f_1(c_i) \cup f_2(c_i) \cup ... \cup f_n(c_i)$ being the result of the concatenation of all identifiers found by all matching functions. Under the assumption that the time difference between nodes is always positive, an edge exists between any two nodes that were created at times t_a and t_b (with $t_a < t_b$), and share a unique subset Λ_z of all their individual cascade identifiers. We furthermore assume that Λ_z and none of the subsets of Λ_z is part of the identifiers found for any node v_c in the set V' containing all nodes that were created at any time t_c with $t_a < t_c < t_b$.

A node that contains a cascade identifier that was not detected for any other nodes before is called the *identifier root*. Similarly, a node without any incoming edges, we refer to as a *network root*, and a node with no outgoing edges a *stub*. Our cascade model, thus clearly yields different structures depending on both the data at hand (e.g. determined by the extent of the Web crawl), and the matching functions applied, which serve to generate the particular cascade identifiers (e.g. based on hashtags, URIs, quotes, keywords, images, or even semantic features such as sentiment), as depicted in Figure 1.

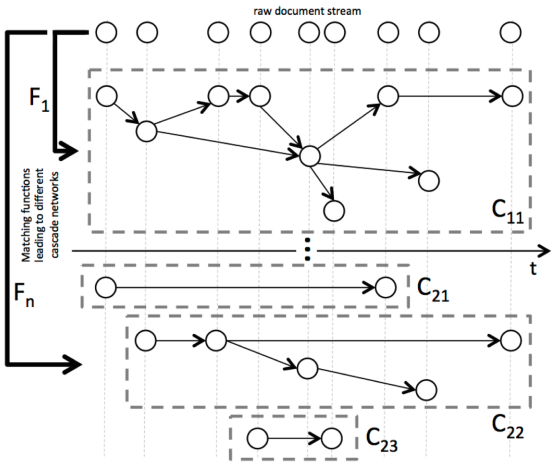

Fig. 1. Depending on the applied matching functions, different transcendental information cascade representations can be generated for the same input data.

To perform a topological analysis of a bounded set of motifs in transcendental information cascades we define 15 primitive node types depending the kind of incoming and outgoing edges as shown in Figure 2. We differentiate edges that are straight, indicating an exact match of the entire set of identifying patterns for two consecutive nodes, and those that branch/merge, reflecting a match of only a subset of the identifying patterns. Node type 1 in this model, for example, means that a node has no preceding node, and is followed by a node that features exactly the same set of identifying patterns like this one (e.g. if in a node we find the hashtags #A and #B, then we find the same two hashtags in the linked consecutive node). A node of type 2 is different from type 1 because the node the edge targets at features only a subset of all identifiers of the node before (e.g. with reference to our example before the consecutive node now only features hashtag #A or #B).

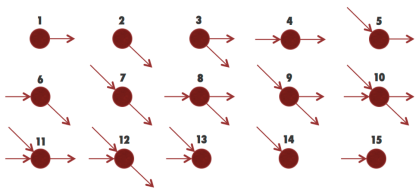

Fig. 2. Overview of distinct nodes types in our cascade model

IV. Experimental Setup

A. Research data

For our study, we use a dataset that contains the contents of the discussion forums from the Planet Hunters[2] project, which is part of the Zooniverse citizen science platform. The dataset consists of $427,917$ posts that have been contributed by participants between the time period of $15/12/2010$ and $16/07/2013$. The Planet Hunters forums support regular forum posts, unlimited in length, and microposts, which are limited to 140 characters and account for more than 90% of all content elements [19].

The use of four particular content matching patterns, which are based upon the characteristics of the Zooniverse platform, form the foundation for our study. We use an exact string matching approach, applying pre-defined regular expressions as done in previous work [2]. The first pattern, *A1*, is based upon the use of hashtags within posts made by users. *Hashtags* are natively supported by the Zooniverse platform as a way for participants to easily group posts with similar topics, just as commonly done on many social media and microblogging platforms such as Twitter. This means that whenever a "#" is directly prepended to a character sequence without whitespace, a link to a list of all posts using this identifier is created. Participants in Planet Hunters make extensive use of hashtags, which are either invented by themselves or recommended by the science team running the project. The second pattern, *A2*, is used to match content that refers to specific object identifiers related to the images shown in Planet Hunters. Over the course of the Planet Hunters project, participants started to link posts to related objects they were presented with during task completion by explicitly including their identifiers. These identifiers feature a consistent pattern of the form "APH[0-9]*", so we developed a second string matching algorithm to use these patterns as cascade identifiers. Pattern *A3* is related to another type of identifier used by the Planet Hunter community to refer to objects in external datasets. In the process of collaboratively identifying exoplanets, some Planet Hunter community members started using Kepler IDs (a pattern matching "KID[0-9]*"), which are codes that uniquely identify exoplanets and other celestial bodies in standard astrononomical indices, to hypothesise correspondences between transits seen in light curves and corresponding exoplanets in these databases. Analogous to the matching of internal object identifiers, we implemented a string matching algorithm for KID patterns. Finally we selected a pattern matcher *A4* that matches all HTTP URIs.

B. Methods for the Analysis of Cascade Properties

Based on the research data and the cascade types, we derived four different cascade networks, one for each of the four types of matching functions (A1, A2, A3, and A4) just described, each applied independently. For each of the four datasets, we then computed a number of measures representative of basic structural properties of the resulting cascades. In particular, we computed the number of nodes, edges, cascades, roots and stubs as well as the cascades sizes. In addition to this, we investigated two measures that are meant to further expand our understanding of the relevance of more complex

[2]http://planethunters.org

	A1	A2	A3	A4
Nodes	66,616	11,061	16,380	24,412
Links	73,595	8,510	9,060	46,842
Cascades	185	1,707	3,233	2,074
Identifier roots	1,458	10,045	14,105	32,922
Network roots	404	3.924	1.901	2.504
Stubs	406	3.899	2.426	2,377
Nodes without any links	539	3,054	4,416	14,014
Avg. Links-per-Node	1.11	0.77	0.55	1.92

TABLE I. OVERVIEW OF THE FOUR CASCADE NETWORKS

and potentially reoccurring sub-structures. First, we determine the Wiener index for each cascade, as proposed in [20]. We refer to the original definition of the Wiener index in [21], given as $W = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}$. Analysing the Wiener index with reference to cascade size allows us to understand whether particular cascades grow in depth or breadth. Second, we compared the constructed cascades in the four datasets with reference to the 15 single node types shown in Figure 2. To complement this structural analysis, our final method focused on informational properties of these cascades. We compute the *identifier entropy* for all our cascades. Identifier entropy is defined as the Shannon Entropy $H(X) = -\sum_{i=0}^{N-1} p_i log_2 p_i$, where p_i is the probability of an identifier occurring in nodes of a particular cascade it has been found at least once within.

V. RESULTS

Table I shows an overview of the four generated cascade networks, based on the four cascade pattern matching methods A1, A2, A3, and A4. These basic statistics show, while the number of nodes in hashtag cascades is significantly higher than when using other identifiers, the number of cascades formed by hashtags (A1) is much lower. This seems to differentiate hashtags from all other three chosen identifiers. However, the average links per node puts A2 and A3 into a joint category (with values of 0.77 and 0.55 respectively) but is by far the largest for URI cascades (A4) with almost two links per node (1.92). Computing the roots and stubs, we find equally both cases, systems with more network roots than stubs (A2, A4) as well as the opposite (A1, A3). The latter case means that more sub-paths of cascades are absorbed over time by merges, than new paths are created by the introduction of new cascade identifiers. Since there were significantly more identifier roots than network roots for all four datasets, it is common for new identifiers to be introduced over the course of any particular cascade in each.

Taking these statistics into consideration, hashtag (A1) and URI (A4) cascades, which are characterised by a high number of nodes, links, and links per node in proportion to the number of cascades and roots, feature a structure that contains many nodes that combine multiple identifiers. The result are many merging and dividing cascade branches. In comparison to this, cascades derived by A2 and A3 tend to feature an initial root node with many outward links (e.g. a initial post with multiple identifiers), which then subsequently diverge and form their own individual cascades without merging again.

In order to consider the structural properties of our four cascade types with respect to existing measures of virality, we computed the Wiener index. Row 1 of Figure 3 shows the value of the Wiener index plotted against the number of links in a cascade. The graphs clearly show that the A2, A3, and A4 cascades resulted in a more sporadic result, with lower virality for larger cascades. In contrast, the virality of the hashtag cascades (A1) constantly increases with the size of the cascade.

The node type statistics show the specific characteristic of the hashtag cascades (A1). Not only that a proportionally low number of network roots and stubs was identified, but the overall amount of nodes with outgoing edges that merge or branch out (types 5 and 6) is significantly higher than when other matching functions are applied (A2, A3 and A4). However, the overall dominant node type is the straight path (type 4), which correlates with the observation of an exponentially growing Wiener index, because it is more likely that cascades grow in depth. It is noteworthy to highlight the specific role of node type 10 in the A4 cascades; this node type indicates that the exactly identifier set recurred intermittently.

Finally, our measurements of the identifier entropy reveal a specific role of hashtag cascades again, and similar profiles for the other three matching functions. In particular, it is shown that most of the hashtag cascades feature even information distributions, which is most likely because they only contain one, or a small number, of identifiers, all with equal probability. Very large hashtag cascades, in contrast, become very random, meaning that even though many identifiers may be included (indicating many informational items being represented in the cascade) the informativeness of the entire cascade overall is minimal. The other three entropy distribution profiles, however, show that there is a more even distribution of information in non-trivial cascades with multiple identifiers, with the largest cascades still falling into the same category as the largest, hashtag cascade.

VI. DISCUSSION

A. Summary of Experiments

Our experiments show that it is possible to generate structurally different cascades from a single source dataset, depending on the pattern matching used. By exploring sub-structures within each of the four resulting cascade networks, we found that in comparison to cascades that use actual object identifiers (KID, APH, URIs), cascades which are based on hashtags tend to be either trivial (single identifier cascades) or consist of multiple roots that are merging and diverging so that they form one massive connected component.

For instance, in A1 cascades, there may be two hashtags, #A and #B, which originate in different, independent posts, by different users. However, over the course of the evolution of the cascade, these hashtags merge, most likely as a consequence of a user bringing them together in a single post. These hashtags then may become part of several merges and diverges, which can end up located within a single stub. As a consequence of this, information can be perceived as lost, as they do not remain present in a distinct cascade, but are subsumed by another. This is reflected in row 2 in Figure 3, where a large proportion of the node types are those that represent merging or diverging cascade paths.

In contrast, the results of cascade types A2 and A3 reveal cascades which are less structurally viral (i.e. exhibit a lower

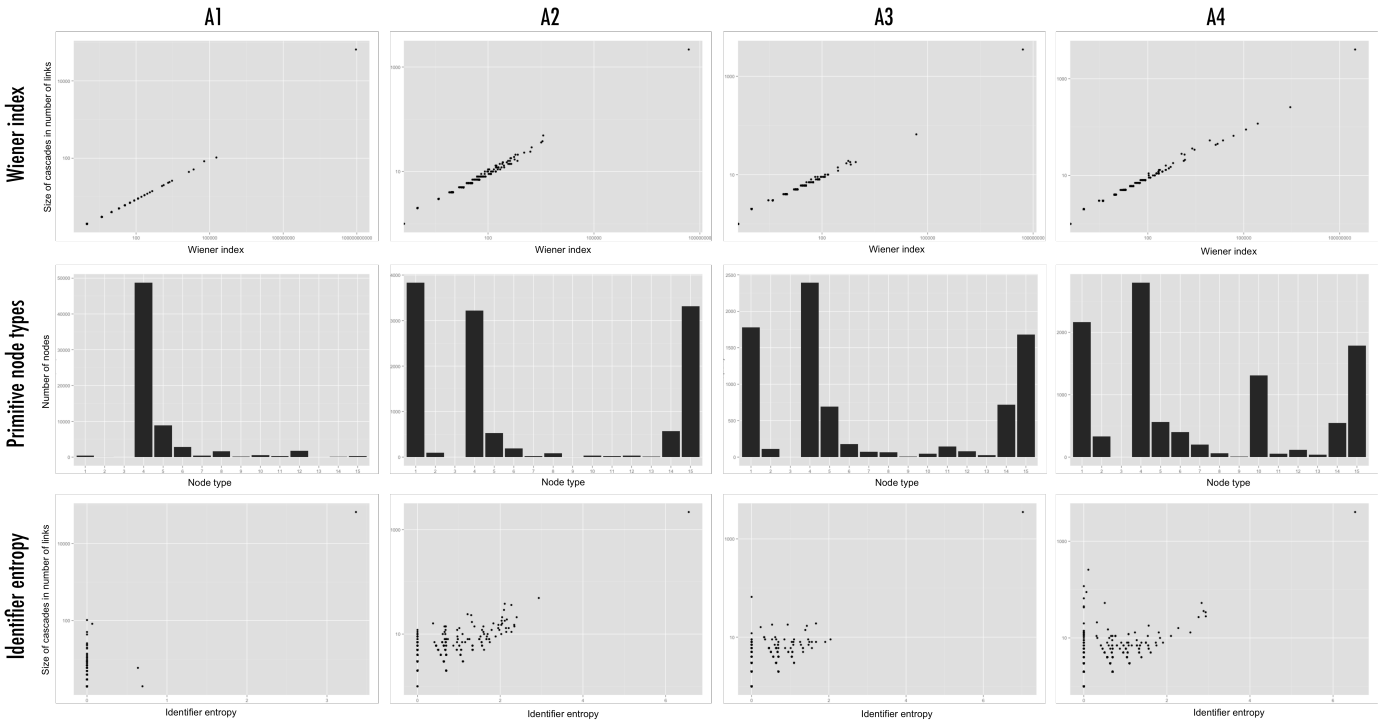

Fig. 3. Overview of the results of the cascade comparison. Wiener index is plotted on a log-log scale; identifier entropy is plotted with a log scale on the y-axis.

Wiener index), thus tending to form shorter chains of single or fewer identifier cascades. As a consequence, information is rarely lost or gained as cascades do not merge often. It is more likely, in such cascades, that when a branch node is observed (for instance node type 3 in 2) it is the root node containing multiple identifiers, which will subsequently form their own sub-cascades. Such cascades will likely branch at the root node, and remain isolated.

B. Structural Properties of Transcendental Cascades

Central to this investigation was the use of different *string matching patterns* in order to understand how the captured information flows differently. In the context of our dataset, which contains discussions of citizen scientists, the examined patterns were based on domain-specific internal and external identifiers of objects of interest as well as hashtags. The analysis of the different patterns revealed varying structural properties when different matching methods are applied. Most significant was the difference of hashtags versus the other object-driven types of identifiers. We found that unlike hashtag cascades, object identifier cascades (A2, A3) tended to interact less with each other. URI cascades were an exception from this rule, as they could be considered object identifier cascades, which featured particular identifier sets that were seen intermittently. In our findings, A2 and A3 cascades had structural characteristics which featured long chains of consistent occurrences of the same identifier, rather than interconnecting and diverging interactions.

This raises questions about the specific nature and inherent features of an entity such as a hashtag, which leads to the observed structural and informational differences to other identifiers. We speculate that this is a result of the socially created conceptual knowledge embodied in hashtag folksonomies, which varies from too generic (e.g. help) to very specific (e.g. ecplipsingbinary) without any boundary between theses two ontological spheres. We suggests that such *collapses* of the informativeness of cascades can be used to reinforce the respective matching functions by contextual refinement (e.g. only treat hashtags as similar information when additional inherent properties of the resources are similar or equal) to derive meaningful results.

Informed by our analysis, we believe that there is large value in not only studying the sub-structures that emerge from transcendental information cascades, but even more by exploring the repetitive patterns of sub-structures which seem to form motifs. This differs from existing research [2]. Rather than examining just the time difference between individual content elements, studying burstiness with respect to repetitive motifs seems to be promising for investigating phenomena like topic promotion, drift or oppression over time. This can become an important means to understand the evolution of online campaigns, and virtual mass coordination and mobilisation strategies from a macroscopic viewpoint, independent from preconceived social determinants.

C. Informational Properties of Transcendental Cascades

The analysis of the experiments raised questions about information gain and loss, and if information cascades are a way to observe and measure this. In our experiments, we found that depending on the pattern matched, there are varying distributions of roots and stubs. In some cases, we see more roots than stubs, which suggests that information that goes into cascades as distinct input, does not come out. This means there is information loss or information gets absorbed when

a particular hashtag wins over others on a particular topic (convergence) for example.

We suggest that an information cascade can be considered as an entity that flows through the Web, channelling and preserving information across time. It therefore has storage and transfer capacity, and as a result is an important aid particularly for distributed communities with few communally-created information storage facilities capable of allowing access to information in a timely manner at the point at which it is needed. Some, but not all, input signals become output signals, so a body of information can evolve over time. Information loss may correspond to information ceasing to be current, or alternatively, a cascade might branch to create divergent cascades whose combined capacity may make up for apparent local losses.

VII. Conclusions and Future Work

In this paper, we took a further step towards an alternative theory for information diffusion online. Rather than focusing to prove conditionality with probabilistic methods, we examined the suitability of using simple, common properties of shared content. Such a *transcendental information cascade* is thus the results of content that is set out within a specific information architecture and the technical capabilities at hand to analyse that content for patterns of similarity. Our experiments focused on the characteristics of different measures to assess the structural and informational properties of transcendental information cascades. In particular, we have shown how the Wiener index, the identifier entropy, and the distribution of particular single node motifs feature varying patterns when different matching functions are applied to the same content. We find significant differences when hashtags are used as cascade identifiers, compared with other content features. We also explain apparent local information losses in cascades in terms of information obsolescence and cascade divergence; e.g., when a cascade branches into multiple, divergent cascades with combined capacity equal to the original. With our study we inform any future application of the transcendental information cascade approach by outlining a basic analytical framework as well as baselines for the interpretation of analysis results.

As part of our future work, we will focus on formalising the interconnections on all levels of transcendental information cascades as represented in Figure 1: (1) nodes and edges in a cascade, (2) full cascades, (3) the system of all cascades derived by applying a particular set of matching functions, and (4) the relationship of any cascade system to all possible other cascade systems representing the same input sequence from a different viewpoint. We ultimately seek to determine which cascade systems are the outcome of (potentially unintended) sequential collective actions. We will approach this by modeling the observed cascades as stochastic processes, and testing the plausibility of different hypotheses about underlying behavioural processes.

Acknowledgment

This work was supported by the EPSRC Theory and Practice of Social Machines Programme Grant, EP/J017728/1.

References

[1] J. Cheng *et al.*, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 2014, pp. 925–936.

[2] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[3] M. Luczak-Roesch, R. Tinati, and N. Shadbolt, "When resources collide: Towards a theory of coincidence in information spaces," in *Proceedings of the 24th International Conference on World Wide Web Companion*, 2015, pp. 1137–1142.

[4] Q. Mei and C. Zhai, "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*. ACM, 2005, pp. 198–207.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," *World Wide Web*, vol. 8, no. 2, pp. 159–178, 2005.

[6] S. A. Myers and J. Leskovec, "The bursty dynamics of the twitter information network," in *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 2014, pp. 913–924.

[7] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[8] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose, "Implicit structure and the dynamics of blogspace," in *Workshop on the weblogging ecosystem*, vol. 13, no. 1, 2004, pp. 16 989–16 995.

[9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 491–501. [Online]. Available: http://doi.acm.org/10.1145/988672.988739

[10] E. Adar and L. Adamic, "Tracking information epidemics in blogspace," in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, Sept 2005, pp. 207–214.

[11] Q. Qu, S. Liu, C. S. Jensen, F. Zhu, and C. Faloutsos, "Interestingness-driven diffusion process summarization in dynamic networks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 597–613.

[12] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs." in *SDM*, vol. 7. SIAM, 2007, pp. 551–556.

[13] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 497–506. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557077

[14] E. Bakshy *et al.*, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 65–74.

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

[16] L. Von Ahn, "Human computation," in *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*. IEEE, 2009, pp. 418–419.

[17] P. Minder and A. Bernstein, "Crowdlang-first steps towards programmable human computers for general computation." in *Human Computation*, 2011.

[18] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 1301–1318.

[19] M. Luczak-Rösch *et al.*, "Why won't aliens talk to us? Content and community dynamics in online citizen science," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[20] S. Goel, A. Anderson, J. Hofman, and D. Watts, "The structural virality of online diffusion," 2013, preprint.

[21] H. Wiener, "Structural determination of paraffin boiling points," *Journal of the American Chemical Society*, vol. 69, no. 1, pp. 17–20, 1947.