# Finding Structure in Wikipedia Edit Activity: An Information Cascade Approach

Ramine Tinati, Markus Luczak-Roesch, Wendy Hall
University of Southampton
Web and Internet Science
{r.tinati,mlr1m12,wh}@soton.ac.uk

## ABSTRACT

This paper documents a study of the real-time Wikipedia edit stream containing over 6 million edits on 1.5 million English Wikipedia articles, during 2015. We focus on answering questions related to identification and use of information cascades between Wikipedia articles, based on author editing activity. Our findings show that by constructing information cascades between Wikipedia articles using editing activity, we are able to construct an alternative linking structure in comparison to the embedded links within a Wikipedia page. This alternative article hyperlink structure was found to be relevant in topic, and timely in relation to external global events (e.g., political activity). Based on our analysis, we contextualise the findings against areas of interest such as events detection, vandalism, edit wars, and editing behaviour.

## 1. INTRODUCTION

Wikipedia is one of the most popular nebula of activity on the Web. In its most rudimentary form, Wikipedia represents a network of human-curated, moderated, and maintained Web pages, which overtime, have become the largest crowd-sourced encyclopedia in existence. Wikipedia has evolved beyond its initial scope and scale, expanding from an initially small group of expert and non-expert editors, into a rich ecosystem, supported by thousands of editors, and champions of diversely motivated volunteers[3].

For Wikipedia, a core challenge is to discover and in certain situations support, emergent phenomena effectively within the vast amount of user and machine generated data. Every second, hundreds of articles are created and revised, edits are overwritten or reverted, abuses are reported, and discussions take place. This stream of activity represents the digital traces of collective human interaction, and studying these streams promises to provide insight into the underlying social activities of such a system. Each seemingly random and uncoordinated entry contains rich information of information, from what text was added or edited, to the interactions and discussions between editors. Whilst individually they represent a single unit of activity (the efforts of an individual), observing them as a stream of activities may reveal something more intriguing.

In this paper we study information cascades with respect to the article edit activity of Wikipedia. Our study draws on 3 months of Wikipedia edit activity collected between January and March 2015, containing $6,051,311$ revisions in $1,572,711$ English articles. We are addressing the following research questions: **[RQ1.] Structure:** What article link structure exists between articles by applying the Information Cascades model to the Wikipedia editing activity? **[RQ2.] Relevance:** Do the information cascades provide context in terms of the content they link, and how does this compare to the structural properties be of the Wikipedia article network? **[RQ3.] Collaboration, Coincidence, and Collective Action:** Can the properties of the constructed information cascades based on the Wikipedia edit stream provide insight into how the seemingly random stream of Wikipedia revisions be used to understand collaboration and collective action?

Our findings show that applying an information cascade model to the Wikipedia edit stream can derive information cascades a network of articles, linked together by the shared identifier found within the edit revision text. In comparison with the explicit Wikipedia article network, the cascade network forms a network of related articles, which is not available within the explicit Wikipedia article network. We also compare the relevance of linked articles using the DBpedia article category labels. By calculating the co-occurrence of topics between articles, we show articles with a cascade are similar in content. Based on these findings, we consider the application of our approach for understanding the characteristics of Wikipedia, and how information cascades could be used as a method to detect internal and external activities, including editing wars, vandalism, detect breaking events, or as a method to search between articles.

## 2. RELATED WORK

The availability of Wikipedia data has inspired scholars to study properties, including structure, contributors, and entities[7, 27], to the the contributions quality of volunteers and articles [18]. Understanding the embedded social practises in Wikipedia has exposed the social norms and practises of Wikipedia, including what motivations editors, what sparks editing wars, and the barriers to entry [25, 15]. There has been particular interest towards understanding article production, and the disputes and editing wars that occur during this process [21], and how measures of quality can be derived to ensure articles are true to the fact [9]. Whilst not directly addressed as a platform to explore information cascades, studies concerning the collaboration between users, and evolution of discussion activity as having features which resemble information cascades [28, 12]. Our work builds upon these studies and explores information cascades that reside in the co-occurrence of entities between Wikipedia edits. Our work aims to apply an information cascade modelling techniques in order to ask questions about the structure, content and social processes which emerges, and based on these findings, we wish to consider what insight can be drawn from this, both from a design perspective, and the implications for a wider community of interest.

Information cascades have been used to model a variety of information sharing practises 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 [2, 13, 1]. Cascades are typically modelled as a dynamic network [26], a directed overlay network on top of a sub-network that represent structures of explicit relationships between entities along which information may diffuse [20, 19, 4, 8].


*WWW'16 Companion,* April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891110.

Transcendental Information Cascades, in contrast, are based on the assumption that there is a natural information flow on the World Wide Web that is not necessarily conditioned by any pre-existing contextual structure. This allows for tracing patterns of activities of human collectives where any pre-defined coordination does not exist or is patchy (e.g. the accumulated editing activity of registered and anonymous contributors on Wikipedia). Kleinberg's work [16] on activity bursts has significantly influenced research studying the temporal properties of human-generated digital content (e.g. [17]), and has also been related to studies of human behaviour at scale [5]. Bursts have become an accepted indicator for the appearance of a topic [24] and can be used to infer meaning by analysing the content in documents that belong to a particular burst. We expand upon this approach; and develop a model which focuses on the frequency of individual information occurring in document streams, and seeks to understand the role of bursts along multiple axes in branching and merging cascades of information co-occurrence.

# 3. EXPERIMENTAL SETUP

Our study uses Wikipedia editing data collected during the time period of 1st January 2015 till the 31st March 2015. The data contains the editing activity of the English Wikipedia article base, collected in real-time from the Wikipedia IRC edit channel, #en.wikipedia. The dataset contains a total of $6,051,311$ Edits, on $1,572,71$ unique articles. For each activity entry we collected the Wikipedia article name, the timestamp the activity was made, the body of the revision (text), and the user associated with the activity. If user is logged in then the Wikipedia registered username will be recorded, if not, we record IP address associated with the edit. The edit text was obtained by using the Wikipedia revision API[1], the query result contains the complete text associated with the revision made.

## 3.1 Methods

**Information Cascade Construction.** We derive our information cascade model based on the model introduced by Luczak-Roesch *et al.* [22, 23]. The method constructs a directed network representation out of selected resources from a discrete time resource stream. Resources are selected to be part of the overall cascade network when a *matching function* matches one or multiple unique informational patterns within the resources' content. Edges are introduced between any two nodes that share a unique subset of all the informational patterns that were matched within their contents and no resource with any of these has been created in the time between the two nodes. In order to construct information cascades using the Wikipedia editing activity dataset, we use a string matching approach [16]. In our case, we apply the string matching function to the text associated with each revision entry. As the text is a complete log of the revision made which includes Wikipedia markup language, and formatting tags, our matching function uses a regular expression to identify nouns phrases within the text. e.g., a matching function with a trigram noun phrase would match 'The White House', 'Barack Hussein Obama II ' or 'The Empire State Building'. Nouns were chosen as they are often used in NLP and entity-extraction techniques, and more specifically, nouns and noun phrases in Wikipedia are often used for entity recognition [10]. We match for all noun phrases identified within the revision text, which means that the cascade root (the edit activity) can spawn an arbitrary number of cascades. As a measure of a cascade's structure, convergence, and divergence, we calculate the Wiener index for each cascade [11]. We also compute the *identifier entropy* for all our cascades, which is defined by 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.

**Methods for Comparison with the Wikipedia Graph.** To compare the linking structure of the cascade article network the Wikipedia article network, we used a reference dataset collected from Wikipedia which contains a complete copy of English Wikipedia article (by their URL), and the link structure between articles[2]. The reference dataset contained $5,716,808$ English article names, with over 130 million article-to-article links. We extract all outbound links (*can_targets*) for a Wikipedia article (*can_source*), for all constructed information cascade pathways, forming the Cascade Article Network (CAN). Using the nodes (article URLs) from the CAN, we then extract the matching pages found within the Wikipedia Article Network (WAN). The match is performed on the article URL (e.g. "en.wikipedia.org/wiki/WWW") Finally, for each *can_source* article, we compare the*can_targets* with the *wan_targets* of the matching article.

**Preliminaries.** We performed experiments in order to determine a suitable noun-phrase length, which included uni-grams, bigrams, and trigrams (or greater). The experiment was conducted on a subset of 100,000 Wikipedia edits, extracted during January 2015. We first experimented with matching all nouns (one or more) within the revision text, however this yielded results which were not computable (or computable in a reasonable duration of time) due to the number of matches between articles. We therefore focused the remaining preliminary experiments on a comparison of using bigram and trigram noun-phrases. We found a large proportion of the bigrams cascades matched articles based on the Wikipedia Markup tags (e.g. 'Line One'). Our manual examination found that the subset of trigram cascades appeared within the bigram cascades (e.g. 'The White House' was found in the bigrams as "The White" and "White House" in two separate cascades). Based on these preliminary results, we decided to use a matching function of a trigram (or greater) noun-phrase.

| Metric | Trigram MF |
|---|---|
| Nodes | 18,896 |
| Links | 17,004 |
| Matched identifiers | 1,745 |
| Identifier roots | 1,599 |
| Stubs | 1,645 |
| Nodes without any links | 146 |
| Avg identifier path length | 11.53 |
| Shortest path (links) | 2 |
| Longest path (links) | 1373 |
| Average path duration (hours) | 369 |
| Longest path duration (hours) | 2133 (88 days) |
| Shortest path duration (hours) | 0 |
| Cascades | 1,379 |
| Largest cascade (links) | 8068 |
| Smallest cascade (links) | 2 |
| Average cascade size (links) | 13.70 |

Table 1: Results of the experiments. The Trigram MF matches on a 3 noun-phrase sequence.

# 4. RESULTS

**Cascades Structure and Properties.** We constructed $1,745$ path cascades from the 3 months of Wikipedia editing data. The length of the cascades exhibit characteristics of a power law distribution; a significant proportion of the cascades were short in path length. Similarly, the distribution of the duration (hours) of the information cascades suggest that most editing activity happens within bursty periods of activity. Furthermore, comparing the duration (hours) of the cascades in respect to the path length, we found that the longer the path, the longer the duration of a cascade. However, certain cascades propagate in a short space of time, relative to the other cascades, which may be an indicator of specific types of activity or editing behaviour.

Assessing the cascade wiener index, and analysing it for all constructed cascades with reference to their size (size in terms of the

[1]Wikipedia Revision API `https://www.mediawiki.org/wiki/API:Revisions`

[2]This dataset can be found via Wikimedia: `https://dumps.wikimedia.org/enwiki/`

number of links), we found a large fraction of cascades with a Wiener index significantly growing with increasing cascade size. This is an indicator for the presence of very long uniform paths without any branches or merges. However, we also detected several cascades following a pattern of a proportionally small Wiener index. This reflects the existence of some more densely connected cascades where more recent nodes are still directly connected to nodes earlier in the cascade. Or in other words, some identifiers occur intermittently with longer periods of no occurrence at all where other identifiers are dominant.

The observations of the structural analysis suggest that different identifiers may play different roles in the cascade and also may contribute a different degree of information to a cascade. We investigated this in more detail by assessing the identifier entropy. We find two cascade types: (a) a significant proportion of cascades with an identifier entropy of 0; (b) the entropy for all captured cascades is lower than 5. While (a) reflects the existence of a significant number of single identifier cascades again, (b) lets us conclude that multi-identifier cascades tend to be dominated by some identifiers resulting in an unequal distribution of the identifiers in those cascades. Both observations support the findings from the analysis of the wiener index in relation to cascade size.

**Burstiness.** We measured three kinds of burstiness: (1) the burstiness of all captured edits independent from the cascade they belong to; (2) the burstiness of all edits captured within specific fully-connected cascade networks; (3) the burstiness of all edits that match a particular identifier (identifier burstiness). As described in Section 2, burstiness refers to periods of high activity in a stream of activity, and offers a way to detect behavior that is correlated with a particular event. In the context of the Wikipedia editing stream bursts of editing activity across a set of Wikipedia articles could be related to some external (or internal) social phenomenon such as a controversial topic, the injection of biased information, or some form of vandalism. The overall burstiness reveals only very few periods of significantly high activity. Naturally, the amount of activity increases as the TIC model will capture additional identifiers the longer the edit stream is observed. This results in an increasing likelihood to match observed edit events to older ones.

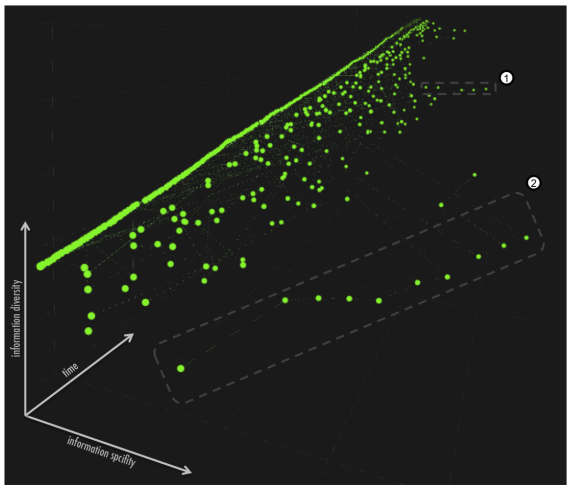



Figure 1: Wikipedia edits in a three dimensional space; (x) time; (y) information diversity as the chronological order in which unique identifier sets are found; (z) information specificity as the index for each unique identifier set which is incremented with each occurrence of the respective set over time.

As a more fine grained indicator of bursts of related information, we computed the cascade burstiness by for each structurally connected cascade network derived from the overall edit stream individually. We observe that it is possible to differentiate between cascades that show a similar burstiness pattern as the overall burstiness and others that are significantly different and become only visible on this microscopic level. TIC allow to map activity streams into a three dimensional space. In Figure 1 we zoomed into a period of 1500 edits happening in about 40 minutes and highlight that within this dense global activity we can identify various local bursts ((1) and (2) mark the most prominent two local bursts). Generally, this mapping of Transcendental Information Cascades allows us to analyse (a) global bursts of high activity involving diverse information and (b) local bursts of significance occurrence of the same information.

**Wikipedia Article Network (WAN) Comparison.** We compared the difference in the link structure of the cascades, and the explicit (embedded) links in a Wikipedia article. We constructed two networks, the Cascade Article Network (CAN), and the Wikipedia Article Network (WAN)[3]. Table 2 provides an overview of the CAN and WAN. For comparative purposes, the metrics of the WAN network have been applied to the sub-set of articles which are contained within the CAN. Figure 2a provides a visual representation of the CAN structure, with three labelled strongly connected components, (A), (B), and (C).

| Metric | CAN | WAN* |
|---|---|---|
| Total Nodes (Articles) | 7,293 | 5,716,808 |
| Total Edges (A-to-A) | 23,560 | 5,705,827 |
| Avg. Edges | 3.1 | 142 |
| Avg. Degree | 6.46 | 343 |

Table 2: A Comparison of the cascade links between articles with the Wikipedia article graph. WAN - Wikipedia Article Network. *The WAN graph metrics are based on the subset of matching Wikipedia articles, not the complete article base.

Due to the articles which reside outside the set of articles identified within the CAN, the WAN has a higher average degree and edges per article. However, in comparison to the WAN's structure which contained one large connected component of articles (within the given subset of articles), the WAN network featured three strongly connected components. As labelled on 2a, these components related to articles containing content about (A) South Korea, (B) the United States of America (Geographic articles), and (C) Political articles.

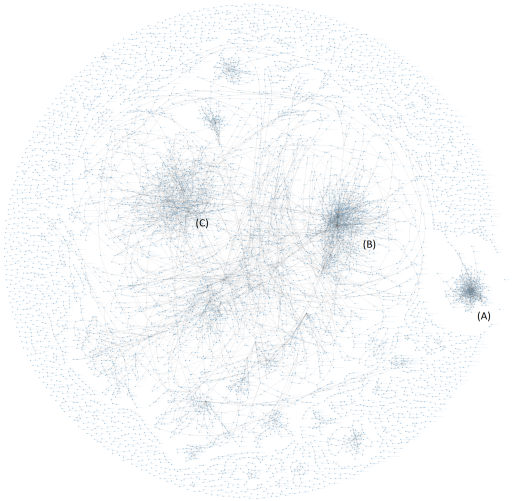
We compared the edges between articles formed by the cascades to the edges within the WAN, and found that only 4.4% of edges in the CAN could be identified within the WAN. Only 2 articles from the CAN had a 100% overlap with the WAN. Furthermore, we found that 94.7% of articles within the CAN had a overlap of less than 1%. These findings suggests that the article links formed within the CAN network may be forming article structure which is not explicitly found within Wikipedia.

| Matching identifier | Associated Root Article | Edges |
|---|---|---|
| U.S. Supreme Court | Hillman v. Maretta | 17,893 |
| NATO Joint Jet | Fighter Pilot | 13,868 |
| U.S. District Court | BJU Press | 5,584 |
| Mehr News Agency | To the Youth in Europe and North America | 2,078 |
| U.S.Religious Landscape Survey | Utah | 1,500 |

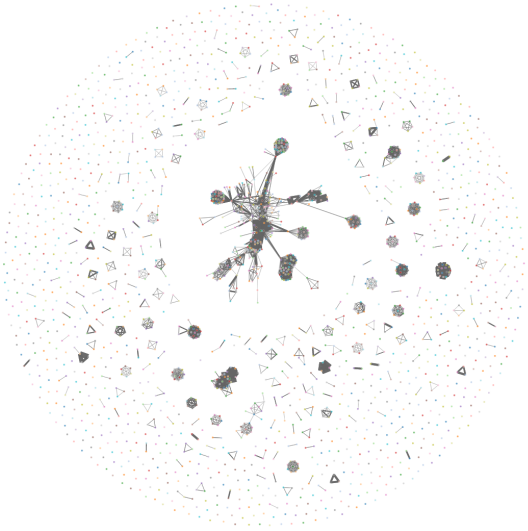
Table 3: 5 highest connected cascades. Each cascade is formed by a particular identifier, and can be associated with a Wikipedia article where the identifier was first used (the root).

**Cascade Category Co-Occurrence.** In order to examine the relatedness between Wikipedia content, we used DBpedia to obtain the category classification labels (*dct:subject*) associated with

[3]A node represent a Wikipedia articles, and an edge represents either a matched identifier between two edits (for CAN), or an explicit link within the Wikipedia graph (for WAN)




(a) Cascade Article Network (CAN): Nodes represent unique Wikipedia articles, edges are shared edits based on a shared identifier matched. A force directed layout has been applied, with edge path lengths determined by edge weight. The strongly connected component (A) contains articles associated with South Korean media, (B) and (C) contain articles related to the USA.



(b) Cascade-to-Cascade path network graph: Nodes are cascades, Edges are the shared articles between cascades. The central strongly connected component is established by the Identifiers shown in Table 3. A force directed layout has been applied, with edge path lengths determined by edge weight.

Figure 2: Article networks

a given Wikipedia article. These labels, which are machine and human generated provide a general classification for the subject (or topic), based on the article's content. We then calculate the co-occurrence of categories between nodes (articles) within a cascade path [14]. Using the co-occurrence measure of a cascade provides us with a way to measure the potential similarity between the subject and content of the articles within a given cascade. Using DBpedia, our queries found, 78.2% of the total articles within the WAN were identified with at least one category. On average, an article was associated with 2 categories. From the 1,745 unique cascades pathways, 521 were found to contain at least one node (article) mapped to a set of categories, and 360 cascades pathways were identified to have two or more articles with categories associated with them. For the analysis, we removed duplicate nodes within a cascade, which were identified as nodes related to the same Wikipedia article, as their categories would be the same, thus skewing the results.

Based on the remaining cascades which had duplicate nodes removed, and two or more nodes with categories associated with them (20% of total cascades), we calculate the co-occurrence of categories between articles within a given cascade. As shown in Table 4, there was an average co-occurrence of 63.6% between article categories within a given cascade pathway. We also extracted the top 10 categories based on co-occurrence frequency. The findings suggest that the articles within a given cascade tend to relate to the same subject or share similar content. We also found that the most frequent co-occurring topics reflect the strongly connected components found in the CAN network, shown in Figure 2a.

| **Metric** | **CTC Network** |
|---|---|
| Total Nodes (Article) | 18,896 |
| Matched Article | 14,776 |
| Unique Categories | 1,605 |
| Avg. Category per article | 2 |
| Avg. Duplicate Article per Cascade | 43.7% |
| Avg. Cascade Category co-occurrence | 63.6% |

Table 4: Overview of the Cascade mapping to DBpedia categories. Avg. Cascade Category Overlap is calculated on cascades with two or more nodes that are associated with different Wikipedia articles

## 5. DISCUSSION

**RQ1: Structural Properties** The structure of Wikipedia can be considered as an *explicit* and *static* network of hyperlinks connecting articles with articles, and with external resources (e.g., hyperlinks to URLs not prefixed by wikipedia.org). We examined whether an underlying structure between Wikipedia articles occurred, and whether this complements, or mimics the explicit linking structure. Our analysis of the wiener index and identifier entropy of the resulting cascades highlights an over-representation of cascades that are long uniform paths with only one matched identifier. Such single identifier cascades can still be suited to find implicit links between articles and detect bursts around trending topics. But it means that only a small proportion of cascades is suited to find implicit relationships between matched identifier phrases.

We conducted the analysis of patterns of burstiness in order to examine the time dimension on the macro and the micro level of the captured edits. The TIC model is based on the principle of capturing elements from a stream that contain a particular informational pattern and bringing subsets of these elements together as branching and merging cascades, when a pattern matches multiple information in some of the elements, so that sequences are linked together. As such it is a generalisation of Kleinberg's approach presented in [16]; based on flat sequences of elements from a stream, only one particular matched information occurs. While the overall burstiness does not show significant bursts from the macroscopic viewpoint of flat sequences of all captured elements, the cascade burstiness reveals bursts not only for trivial single identifier cascades. The identifier burstiness provides further evidence for the

existence of identifiers with correlated bursts. This part of our analysis emphasises the role of the time dimension with respect to the implicit article structure we derive. Some identifiers burst only once or they burst intermittently and then stop occurring. That means that the relationship between the articles could be seen as being of temporary nature or at least of lower relevance the longer the time since the last burst. Consequently, a link foraging function could be a useful component in the cascade article network construction.

**RQ2: Structural Relevance** Unlike the embedded hyperlinks of an article, our constructed cascades represent temporal pathway between articles, which emerges at a given point in time, without explicit linking or conscience action. By comparing the structure of the cascade article network with the Wikipedia article network, we found structural differences within the respective graphs, as well as differences in the associations within articles. As only 4.4% of article links within the cascade article network overlapped with the Wikipedia article network, we assume that the cascades exposed pathways between articles which are not necessarily navigable using the embedded links on an article. The analysis of article *category* co-occurrence also demonstrated the relevance of the cascade pathways. Within the cascades, of those which could be associated with a category (derived from DBpedia), over 60% of the articles shared one or more category, which suggested similarity within the different articles's content. These categories also mapped to articles within cascade article network which were part of three strongly connected components, as show in Figure 2a. To illustrate this further, we consider the cascade formed by the identifier *U.S. Supreme Court* as shown in Table 3. This cascade contained many links, had several bursts of activity, and spanned for over 2 months. and was one of the most connected cascades identified within the Cascade-to-Cascade network. A closer inspection of the links formed within the cascade articles, revealed that the identifier linked together articles which were not explicitly linked together within the Wikipedia article network. In this context, the information cascades provided an alternative pathway, with a different set of related articles – and thus related knowledge – compared to what is shown explicitly within the article content. In this cascade example, the originating root article *Hilland V. Maretta*, contained content about about a decision made by the U.S. Supreme Court, which then linked to an article discussing suppressing the freedom of speech. This forward chaining of articles using the shared identifier subsequently linked another 847 articles related to legal decisions within and outside the United States of America.

The cascade example described before illustrates how the editing activity can provide an individual with a pathway through articles which could not explicitly found with the Wikipedia article network alone. Moreover, this could also be relevant to a specific point in time, where, taking for example the burstiness of a cascade, may indicate a pathway of edits (thus connected articles) which relate to some external event or phenomenon. Building on the findings of RQ1., whilst the given example was restricted to a single identifier path, the TIC model has shown that there are circumstances where cascades merge (and branch, which in this case, would be where two or more identifiers (in our case, *U.S. Supreme Court*, *U.S. District Court*, and *Mehr News Agency* are used within a Wikipedia edit. In this case, the edit (which is related to an article), merges these cascade pathways, which may represent similarities between topics, content, or relevant information.

**RQ3: Collaboration, Coincidence, and Collective Action** Understanding emergent socio-technical behaviour and phenomena at the micro- and macro-level is one of the key drivers for Web Science research [6]. Central to this study was using a cascade model, which has been shown to as a suitable method to expose serendipity and emergent phenomena in similar crowdsourced environments [23]. The construction of cascade identifier paths and information cascades reveals temporal editing pathways between Wikipedia articles, which offer an alternative route for individuals to navigate between articles. The cascade structural properties and burstiness analysis reveals how cascades exhibit different characteristics, based on their temporality, their relevance to external phenomenon, as well as external events. We found that certain cascades exhibit levels of high-activity, and draw together articles of similar content and subject, both within a single identifier pathway, or across multiple identifiers.

In the context of related Wikipedia research, which include areas such vandalism and conflict detection, editing wars, or the identification of trending topic, we consider the application of the TIC model as a suitable method to capture and expose a temporal and dynamic structure between articles. For example, by using this method, we observed a burst of activity which features a series of edits made within a short duration of time. These edits, related by the identifier *U.S. Coast Guard* begin on the Wikipedia article, *Edward Snowden*, and capture a chain of successive edits to the same, and other articles, such as *Hayfield Farm Community* and *Northwest Airlines*. Furthermore, a broader inspection of the timeframe when the cascade first emerged, coincided with a presentation given by Edward Snowden, at an international conference (SXSW). In this example, we observed a relationship between external phenomenon, and the emergence of a short, bursty cascade. Its characteristics could potentially be an indicator for detecting controversial real-world events, or trending topics. Similarly, we are also able to observe more local phenomenon, such as the pathway found using the identifier: *U.S. District Court*. In contrast to the previous example, this cascade extends over a long period of time, and links together the editing debate found in the *Same-sex marriage in the United States* and *List of U.S. state laws on same-sex unions* articles. In this example, we find the re-occurrence of articles within a single pathway, potentially indicating the back-and-forth editing activity between editors. In both examples, we witness how the construction of cascades reveals coincidental collaboration and collective action within a stream of activity which has no a prior structure or connectivity. As the TIC model works as a context-free approach to reveal information cascades, our findings suggest that it has application both as a method for detection and prediction of phenomenon.

# 6. CONCLUSIONS AND FUTURE WORK

In this paper, we applied an information cascade model to a Wikipedia dataset containing over 6 million Wikipedia edits on over 1.5 million English Wikipedia articles. The purpose of the study was to answer questions related to how constructing information cascades based on Wikipedia article editing activity could be used to reveal insights about the structure, and the emergence of micro and macro-level socio-technical phenomenon.

Based on the experiments conducted, our findings have shown that the combination of the TIC model with the stream of Wikipedia results in the formation of dynamically generated cascades. The TIC model has shown to be a suitable method to extract a network of articles, formed by the co-occurrence of entities within the edit stream. Our findings have shown that using information cascades, it is possible to construct links between articles which are similar in content and subject without a priori knowledge about their relationship within the Wikipedia article network. More than 95% of these cascade article links were not present within the links found within a Wikipedia page; thus the cascade model offers alternative, and in some situations, hidden routes for individuals to navigate along. Furthermore, unlike the static *a priori* structure of the Wikipedia article network, that temporal nature of the cascades promotes linking between articles which may coincide and remain active during the burst of some external phenomenon, potentially useful as a mechanism to derive insights beyond the "walls" of Wikipedia.

This study has demonstrated the potential of studying the action of ad-hoc collectives that form around internal and external events (e.g. edit war around same sex marriage vs. Edward Snowden speaking at SXSW) rather than following any *a priori* coordination or planning. Investigating the properties of the cases described in Section 5 lets us suggest that features such as the duration of the bursty period (e.g. the Snowden speech burst was short and dense, while the edit war bursts for much longer) as well as the number of articles linked by the cascades (e.g. the Snowden speech burst

involves many articles while the edit war features a very concise set of articles). Such features could be used for the automatic detection of an internal or external trigger caused the action. Our future work will investigate the validation of our assumptions with ground truth data about edit wars, with the aim to devise a method to detect different kinds of collective actions on Wikipedia by looking at the edit stream. We will also expand on the inherent TIC features used so far and add contextual features such as the editing users involved, including their roles within the Wikipedia community.

# 7. ACKNOWLEDGEMENTS

This work is supported under SOCIAM: The Theory and Practice of Social Machines, funded by the UK EPSRC under grant EP/J017728/2.

# 8. REFERENCES

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

[2] Adar, E., Zhang, L., Adamic, L. A., and Lukose, R. M. Implicit structure and the dynamics of blogspace. In *Workshop on the weblogging ecosystem*, vol. 13 (2004), 16989–16995.

[3] Antin, J. My kind of people?: Perceptions about wikipedia contributors and their motivations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 3411–3420.

[4] Bakshy, E., 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), 65–74.

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

[6] Berners-Lee, T., Weitzner, D. J., Hall, W., O'Hara, K., Shadbolt, N., and Hendler, J. a. A Framework for Web Science. *Foundations and TrendsÂő in Web Science 1*, 1 (2006), 1–130.

[7] Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. Preferential attachment in the growth of social networks: the case of Wikipedia. *Physical Review E 74*, 3 (2006), 4.

[8] Cheng, J., 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), 925–936.

[9] De la Calzada, G., and Dekhtyar, A. On measuring the quality of wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, ACM (New York, NY, USA, 2010), 11–18.

[10] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proc. VLDB Endow. 6*, 11 (Aug. 2013), 1126–1137.

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

[12] Gómez, V., Kappen, H. J., and Kaltenbrunner, A. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, HT '11, ACM (2011), 181–190.

[13] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, ACM (New York, NY, USA, 2004), 491–501.

[14] Ito, M., Nakayama, K., Hara, T., and Nishio, S. Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, ACM (New York, NY, USA, 2008), 817–826.

[15] Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, ACM (New York, NY, USA, 2007), 453–462.

[16] Kleinberg, J. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery 7*, 4 (2003), 373–397.

[17] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. On the bursty evolution of blogspace. *World Wide Web 8*, 2 (2005), 159–178.

[18] Lam, S. T. K., and Riedl, J. Is wikipedia growing a longer tail? In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, ACM (New York, NY, USA, 2009), 105–114.

[19] Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, ACM (New York, NY, USA, 2009), 497–506.

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

[21] Liu, J., and Ram, S. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst. 2*, 2 (July 2011), 11:1–11:23.

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

[23] Luczak-Roesch, M., Tinati, R., Van Kleek, M., and Shadbolt, N. From coincidence to purposeful flow? properties of transcendental information cascades. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (Aug 2015).

[24] Mei, Q., and Zhai, C. 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), 198–207.

[25] Pfeil, U., Zaphiris, P., and Ang, C. S. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication 12*, 1 (2006), 88–113.

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

[27] Stuckman, J., and Purtilo, J. Measuring the wikisphere. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (2009).

[28] Taraborelli, D., and Ciampaglia, G. L. Beyond notability. collective deliberation on content inclusion in wikipedia. In *Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on*, IEEE (2010), 122–125.