LOOM: Showing the Dynamics of Power Laws in Twitter Data

Maryanne Doyle, Mark T. Keane
Insight Centre for Data Analytics & School of Computer Science, University College Dublin
maryanne.doyle@insight-centre.org, mark.keane@ucd.ie

***Abstract*—LOOM is advanced as a new visualisation for changes in ranks and trends in power-law data that is changing dynamically over time. A comparison between LOOM and existing methods for visualising such data (e.g., time-series graphs, typical analytics dashboards). Several exemplar data sets are shown, using LOOM, drawn from the tracking of news stories on Twitter. The basis for the LOOM visualisation is elaborated and it is shown how it avoids the pitfalls arising in other line-graph representations.**

Keywords-exploratory visualisation, visual analytics

I. INTRODUCTION

Many aspects of the natural world and human activity are reflected in power laws, making them a key focus for both the sciences and the humanities [1]. In recent years, they have become even more ubiquitous, reflecting a plethora of small-world effects in cyberspace; e.g., capturing online phenomena [2] from the plotting of trending hashtags, to the linking of web-pages [3], to trending likes on facebook posts. For instance, on Twitter, the hashtags used to reference a given news event are typically distributed as a power law; that is, there is one dominant hashtag that emerges as the most frequently used, with a rapid falling off in frequency to other hashtags, ending in a tail of many infrequently-used hashtags. Increasingly, researchers and practicioners want to track these dynamic changes in the distribution of these items in real time, not just plot them retrospectively.

However, the traditional methods used to graph such distributions are not particularly useful for tracking such dynamic changes (see Figure 1). What is really required is a new representation that is tailored to the specific needs of such users, a representation of the data that surfaces the key aspects within these power law datasets that users want to see[1]. In this paper, we present a proposal for representing such power-lawed data using a method that shows key aspects of the behaviours of interest. To illustrate this method, we apply it to datasets of trending hashtags from Twitter, though these methods are applicable to any power law, or heavy-tailed dataset where changes over time are of interest.

[1] Think of the present representation as being analogous to DET curves that were designed to "re-represent" the data in ROC graphs to allow better comparisons of systems [4].

A. Twitter & Trending Hashtags

Created in 2006 as an online social networking service, Twitter enables users to interact through short messages called tweets. Nowadays Twitter has a worldwide reach with 319 million users sending approximately 500 million tweets per day. Twitter has been branded as the social network for news dissemination [6]; indeed, many journalists and news organizations now expend considerable effort tweeting their news articles, in ways to attract maximal attention and engagement [7, 8]. In attracting attention to one's tweeted news, the use of the right hashtag for a story has become critical [9]; if a journalist does not use the most-commonly-used hashtag for the story then their news is less likely to reach a target audience [10, 11]. Twitter's own TweetDeck application tries to provide broad information on trending hashtags but it is not fine-grained enough to spot the emerging behaviour of competing hashtags on a story. Part of the problem here is that, as a story breaks, there are often several competing hashtags that may take several hours to "fight it out" for popularity [12, 13]. So, journalistic users really need to have a view of this competition, to be able to project forward to what may be the winning hashtag or to know that there are several potential winners in the set, with a view to using them in their news tweet.

The goal of the current work is to present such users with a visualisation – called LOOM – that allows hashtag trending data to be easily and intuitively read. This visualisation needs be able to handle the challenges inherent in these datasets, such as a large scale, a high volume of tweets and the volatile and dynamic changes that occur within this use-case.

B. Visualising Trending Hashtags

There are many existing visualisations that attempt to represent data exhibiting power law behaviour, though few are very dynamic. Figure 1 shows the items within a distribution plotted by frequency on logged axes. This representation shows that the data is distributed as a power law, those items in the approximately linear portion of the graph on the left are in the "head" of the distribution (the most frequently occurring, or most popular items) and those in the flat portion on the right are in the "tail" of the distribution (the least popular items). However, this visualisation does not represent time and as such offers the reader no insight into how the items within the distribution

have changed in frequency over time. One could attempt to represent changes in frequencies by creating a graph for each time-step as a separate frame in an animation, but tracking data-points in this way does not seem very promising.

Time series graphs do include time in their representation of the data (see Figure 2). In Figure 2 global mean precipitation is shown over a period of several decades, with each coloured-line representing a different source of precipitation data. These line graphs are used primarily in the field of statistics and can show trends, seasonality, outliers and discontinuities [14]. However, to a non-expert viewer these representations could be difficult to read, as the jagged paths of each item take up much of the graph space and overlap frequently, decreasing the ease of readability.

Google Trends presents time-series data using values averaged over a time window resulting in smoothened paths for each item graphed (see Figure 3). In Figure 3 we see a plot of searches made using GoogleTrends for Mark Zuckerberg (in blue), Bill Gates (in red) and Enda Kenny (in yellow; Ireland's Taoiseach/Prime Minister). The X-axis shows time (a period of six months) and the Y-axis shows a scale from 0 to 100 representing a measure of search interest relative to the highest point within the focal dataset. The highest point in the graph appears in November 2015 (when Mark Zuckerberg announced he would take two months of paternity leave) and all other Y-values are determined relative to this event. This use of a relative scale is one way to capture a wide range of values in one frame. However, this approach to handling scale is problematic when one series has much lower values (i.e., Enda Kenny in yellow in Figure 3) relative to the Y-axis defining event. Such lower-frequency items tend to flatline, hiding changes that are occuring in them, again decreasing the readability of the data.

Several dashboards exist for the analysis of Twitter data and hashtags such as Keyhole2, Hashtracking3, Tweetreach4 and TweetBinder5. These dashboards offer a variety of services (e.g., monitoring multiple social networks, analysing hashtag sentiment) but they all use line graphs to show changes in hashtags over time, akin to the GoogleTrends solution (e.g., see Figure 4).

So, all of these representations have issues dealing with multiple items in the same frame, especially when the items demand significantly different scales. As such, there is clearly a need for a new representation that can present such data items in a clear manner for both expert and non-expert users. Here, we report such a representation using a modified line graph that better delivers what users need to see. This representation also handles the challenges in scaling and can be used as a visual tool to analyse the behaviour of such datasets.

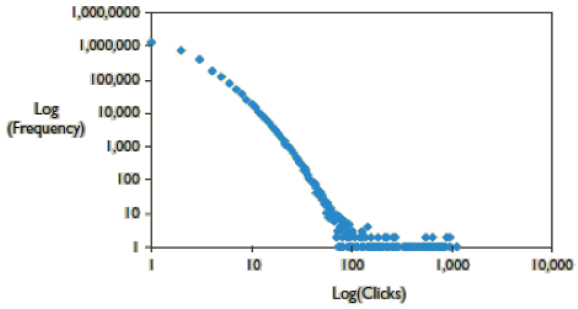

Figure 1. Frequency distribution of mobile web surfing, Halvey et al [5]

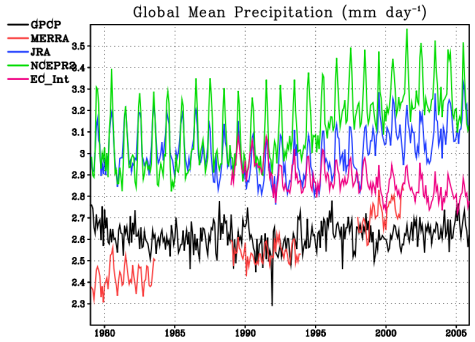

Figure 2. Time series line graph of global precipitation[6]

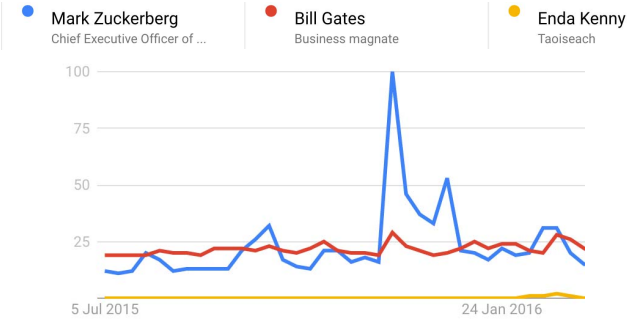

Figure 3. Google Trends line graph

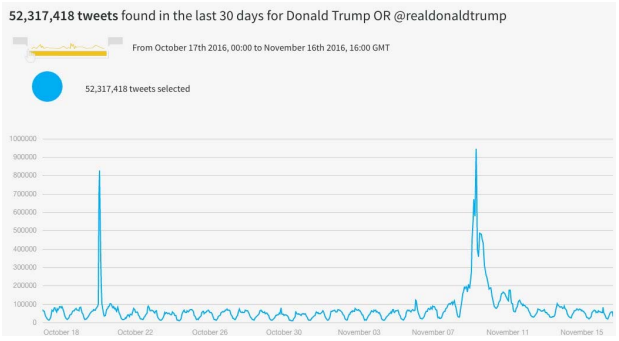

Figure 4. Example of a line graph in analytics dashboard, TweetBinder[7]

2 Keyhole: http://keyhole.co/

3 Hashtracking: https://www.hashtracking.com/

4 Tweetreach: https://tweetreach.com/

5 TweetBinder: https://www.tweetbinder.com/

6 M Bosilovich, "Precipitation Time Series" http://cookbooks.opengrads.org/images/3/3b/Precip_time_series_ss.png

7 TweetBinder, "Mentions to Donald Trump" https://pbs.twimg.com/media/CxZcbJNXgAALZ-P.jpg

II. LOOM: TREND-TRACKING REPRESENTATIONS

The LOOM representation was designed to better handle the problems in tracking trends in datasets with power-lawed behaviours. Specifically, LOOM was designed to (i) be visually smoother than time series plots, (ii) preserve readability across a wide variety of scales and (iii) allow the comparison of the dynamic changes in multiple items in an informative way.

To illustrate how LOOM works, consider a sample dataset of just three hashtags scoring the number of uses each item has in tweets, that change over a focal time period. LOOM applies three transformations to turn the traditional time-series-like graph into a LOOM graph.

A. Smoothing with Cumulative Values

Our sample of raw data for three hashtags shows how the frequencies of the hashtags "a", "b" and "c" change over a 5-hour period (see Figure 5). Hashtag "a" has a high volume of tweets at each hour, "b" has a low volume at each hour and "c" undergoes a significant increase by hour-3.

In LOOM, the first transformation we make to the data is to use cumulative values over time rather than raw values (see Figure 6). This operation smoothens the data encoding the history of an item's progress to its current position, allowing the user to distinguish between a new hashtag undergoing a sudden burst of growth ("c") and an established hashtag ("a") of sustained popularity.

The next step is to normalise the data so that changes in the frequency of one hashtag are represented relative to the set of hashtags as a whole (see Figure 7).

This transformation is achieved by plotting the cumulative count of tweets of a given hashtag over the cumulative sum of all tweets, upto a given point in time (see Formula 1). This normalisation avoids the problem of ever-increasing cumulative values over the total period of time. It also shows how the usage of all three hashtags falls towards the end of the time period in question.

$$freq_{norm} = \frac{\sum_{t=1}^{T} h_i(t)}{\sum_{t=1}^{T} h_{all}(t)} \qquad (1)$$

Next we use a log scale on the Y-axis so that readability is maintained across the graph from the most used hashtag down to the least used. This allows us to see the changes in volume of "b" (see Figure 8).

This new way of coding the data gives us a perspective on changes at the lower end of the scale removing the problem of a visual flat-line that obscures values in other line graphs (e.g., Google Trends, see Figure 4). In this respect, it is interesting to compare Figure 5 with Figure 8. In the final LOOM representation, we can see the change at both the high and low ends of the scale in a single frame. We can also tell the difference between highly ranked hashtags as a result of sustained popularity as opposed to hashtags that undergo a sudden burst in use. In the next section, we show how this representation works when it appled to larger, real-world datasets.

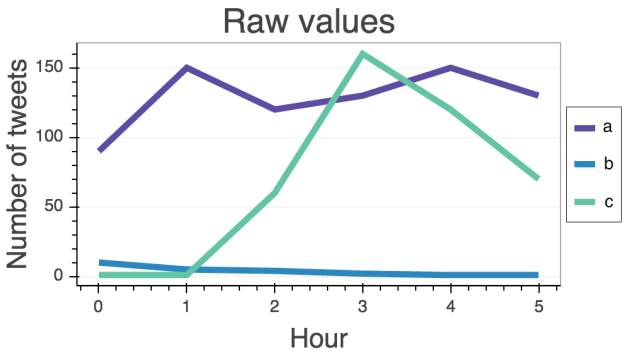

Figure 5. Raw values in LOOM

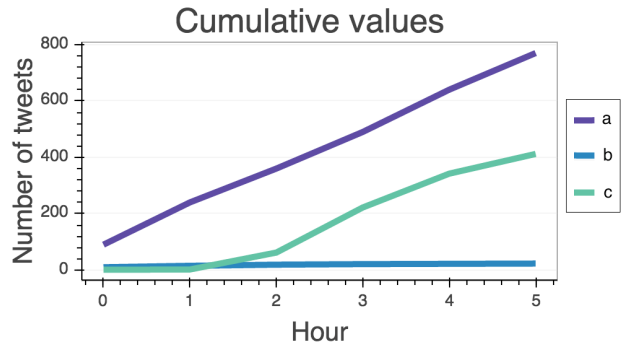

Figure 6. Cumulative values are used

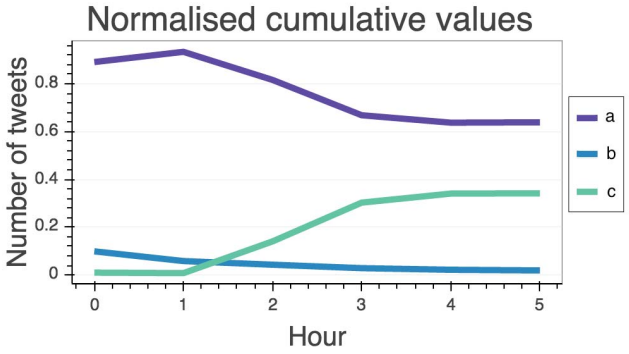

Figure 7. Normalisation is applied to the cumulative values

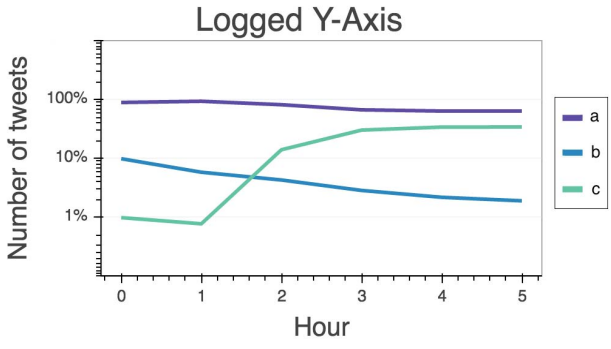

Figure 8. A logged scale is applied to the Y-axis

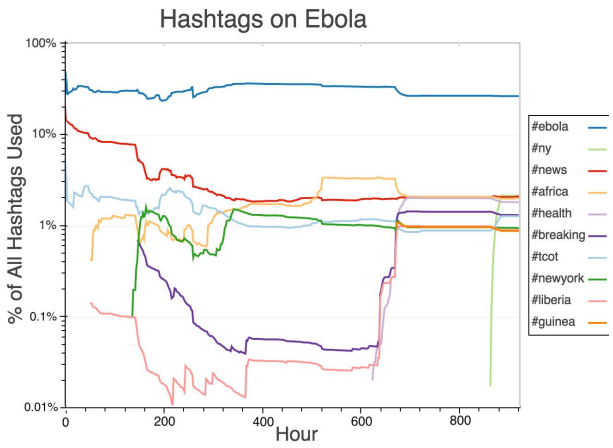

Figure 9. Hashtags on Ebola

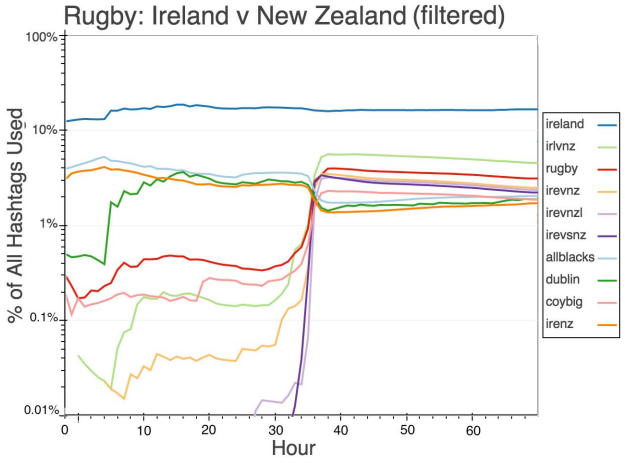

Figure 10. Filtered hashtags on rugby

III. LOOM: The Long and the Short of It

In this paper, the data used was collected using a system -- called Hashtagger [10] -- that extracts trending hashtags from real-time news stories and streamed Twitter data. Hashtagger retrieves tweets about a news story using keywords from article headlines and summaries and outputs the hashtags referencing the story. It also collates counts of the all the hashtags of relevance to a particular breaking news story.

A. Long Stories in LOOM

The first corpus of tweets covered the Ebola outbreak in 2014, involving trending hashtags over a 40-day period beginning on October 10th 2014 and ending on November 19th 2014. This corpus contained 274,793 tweets and 12,306 unique hashtags. This dataset typifies a news story that unfolds over a longer period of time and has many sub-events, a news story as defined in a typology by Zubiaga et al [11].

Figure 9 shows LOOM's visualisation of hashtags about the Ebola story, clearly indicating that #ebola is the highest ranked hashtag positioned at the top of the graph. Also, note that the difference in frequency between it and the next highest-ranked hashtag is large and relatively constant after the first 200 hours. The fact that #ebola maintains its horizontal trajectory shows that its frequency of use per day is consistent relative to the total number of hashtags in the story's 40-day life.

Notably, the hashtag labelled ***#ny*** (shown in the right of the graph in lime-green) has an almost vertical trajectory ascending from the bottom to top of the graph. We call this a "breaker". These are hashtags that show a very rapid burst in frequency, clearly shown in their vertical tracectories as they cross other hashags and ascend in rank out of the tail, into the distribution's head. This breaker (***#ny***) represents a sub-event within the Ebola story that occurred on November 15th 2014 (in Figure 10 between hour-888 and 912) reflecting a story about a doctor who contracted Ebola while working in Sierra Leone, who was flown to the US for treatment.

B. Short Events in LOOM

The second corpus of tweets concerns a story with a much shorter life, with hashtags about an international rugby match between Ireland and New Zealand that attracted significant Twitter attention. This event is in the category of current events as defined in a typology by Zubiaga et al [11]. The match took place on November 24th 2013 between 14:00 and 16:00 and the dataset covers a 3-day period from 00:00:00 the day before the event until 23:59:59 the following day. In this 72-hour period the match fell in hours 38 and 39. The dataset used was also collected using Hashtagger [10] and the results were filtered using a series of keywords about the event (e.g., "rugby", "Ireland", "New Zealand"). This filtering was designed to ensure that as many tweets specifically about the rugby match as possible were included in the corpus. This filtered dataset consisted of 121,884 tweets involving 12,190 unique hashtags.

In Figure 10, we see a high constant use of ***#ireland*** (i.e., Irish team), ***#dublin*** (where the game took place) and ***#all-blacks*** (i.e., the nickname of the New Zealand team). Then around hour-40 we see breakers for ***#irevnz***, ***#irlvnz***, ***#irevsnzl***, three hashtags specifically referring to the match, though we see that, in time, ***#irlvnz*** dominates. Overall, we see a very coherent picture emerge in the evolving hashtag-landscape for this event; indeed, one that requires little specialist expertise to read.

IV. Emergent Properties of LOOM

While a representation, like LOOM, is designed to manifest key properties, a hallmark of good representations is that they will have "emergent" properties (i.e., beneficial features of the representation that are additional to those envisaged in the original design). LOOM appears to have one such emergent property; namely, that it appears to highlight key hashtag behaviours, even in unfiltered tweet collections.

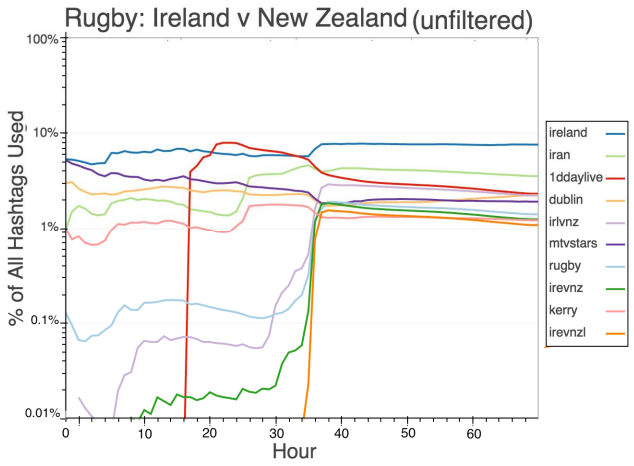

Figure 11. Unfiltered hashtags on rugby

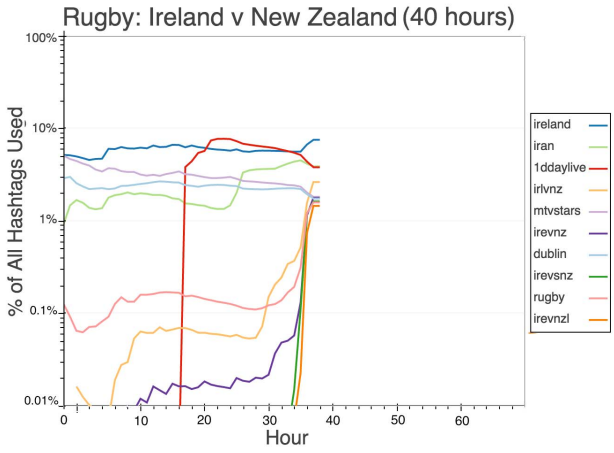

Figure 12. Unfiltered hashtags on rugby until the time of the match

Recall, that the original rugby story corpus was filtered to narrow down the tweets to those about the rugby-match story. We collected a second tweet corpus, during the same 72-hour period as the original set, but this time did not apply any filters. As a result this dataset is much larger (i.e., 288,832 tweets involving 38,113 hashtags) and noisier (i.e., visual inspection of Figure 11 shows that it contains many non-rugby related hashtags).

In Figure 11, we show all the hashtags that occur in this unfiltered set. Here, the graph shows a much noisier dataset in which many different stories are mingling (e.g, about Iran, MTV and an event involving pop band One Direction called "1D-Day"). However, even in this dataset, the hashtags related to the rugby game show some prominence in the ***#ireland*** hashtag and the sudden peak in the ***#irlvnz*** hashtag. Interestingly, LOOM also allows us to explore this much noisier dataset to reveal more of what was going on during the day, without having to filter the data. To explore this further, we re-computed the data with respect to hour-40 (the time at which the match is played).

So, in Figure 12, where we have re-computed the cumulative frequencies up to the 40th hour, we see that we get a different picture of the unfiltered data that reflects a lot of what we see in the filtered dataset. The hashtags ***#ireland***, ***#irlvnz, #rugby, "#irevnz", #irevnzl, #irevsnz*** and ***#dublin*** emerge as being dominant. It is interesting to compare Figure 11 (based on the filtered dataset) and Figure 12 (on the unfiltered set); basically, it shows that the use of cumulative frequencies combined with normalising on the total-tweet-number at a given time-point acts to shake out emerging patterns in the noiser data[8]. This result is quite a significant finding, as it shows that LOOM has an emergent property that allows users to "find" patterns in the data when it is used to track an unfolding event. Critically, this shows that LOOM could be used as a data analytics tool for noisy datasets rather than just a visualisation of "clean" data.

[8] To make the matching that occurs here clearer we manually assigned the colours used in Figure 12 for the hashtags in common with Figure 10.

V. Other LOOM Applications

Applications in the field of news and journalism are an obvious opportunity for LOOM, where the users are journalists, PR-companies or members of the public who want to follow stories as they unfold. Although the examples discussed in this paper are based on data from Twitter, the LOOM representation could be applied to any dataset in which changes occur to data over time in a power law distribution. For example, the data points could just as well be changes in sizes of populations of cities [16], fluctuations in stock prices [17], properties of genomes [18] or changes in word usage across social media platforms where trending behaviour is important.

The re-computing of data up to a key, selected hour (shown in Figure 11) is an operation that would be very useful as an analytics tool, so we are currently developing LOOM as an interactive, visual-analytics tool where the user can select points of interest to re-compute to examine the data (as well as other operations).

VI. Conclusion

In today's world of big-data, where there is a surfit of information to explore, there is an urgent need for visualisations to capture the dynamic behaviour of data over time. In this paper, we present three main novelties:

- an analysis of the key requirements that need to be represented in a visualisation for story tracking, showing how current techniques fail to meet these requirements
- an encoding of information needed to produce a visualisation called LOOM which captures all of the key properties identified and that can scale to large and dynamic datasets
- the dynamic properties of LOOM shown by testing it on real world datasets, currently being analysed in social media (i.e., Twitter hashtags).

In summary, LOOM represents noisy data that changes over time in a clear and understandable manner that is easy to interpret. It can show extremes of scale within a dataset (as is typical of power-law datasets) without losing readability of the highest or lowest frequency items. LOOM has great potential to support research efforts across many disciplines where analysis of datasets with power law distributions that change over time is of interest (e.g., physics, biology, earth and planetary sciences, economics and finance, computer science, demography and the social sciences [19]). Essentially, anywhere where changes to the items within a power law dataset over time are of interest, LOOM has a role to play in exposing and representing what is happening. We see that this broad spectrum of application and efficacy of use means that LOOM should be used as a tool to gain insight to raw and noisy data in many fields.

ACKNOWLEDGMENT

This work was supported by Science Foundation Ireland under grant SFI/12/RC/2289 and grant 07/CE/I1147 (Insight Centre for Data Analytics).

REFERENCES

[1] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review* 51.4 (2009): 661-703.

[2] Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos. *"On power-law relationships of the internet topology."* ACM SIGCOMM computer communication review. Vol. 29. No. 4. ACM, 1999.

[3] Adamic, Lada A., and Bernardo A. Huberman. "Power-law distribution of the world wide web." *science* 287.5461 (2000): 2115-2115.

[4] Martin, Alvin, et al. "*The DET curve in assessment of detection task performance*." National Inst of Standards and Technology Gaithersburg MD, 1997.

[5] M J Halvey, M T Keane, B Smyth *"Exploring social dynamics in online media sharing"* International World Wide Web Conference 2007

[6] M. Osborne and M. Dredze. "*Facebook, twitter and google plus for breaking news: Is there a winner?"* International Conference on Weblogs and Social Media, ICWSM 2014

[7] C Orellana-Rodriguez, D Greene, MT Keane *"Spreading the news: how can journalists gain more engagement for their tweets?"* 8th ACM Conference on Web Science, 107-116

[8] D. L. Lasorsa, S. C. Lewis, and A. E. Holton. "*Journalism Studies*" chapter Normalizing Twitter. 2011.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. *"Measuring user influence in twitter: The million follower fallacy"* ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social, 2010.

[10] B Shi, G Ifrim, N Hurley "Learning-to-rank for real-time high-precision hashtag recommendation for streaming news" 25th International World Wide Web Conference

[11] Zubiaga, Arkaitz, et al. "Towards real-time summarization of scheduled events from twitter streams." *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 2012.

[12] K Lee , J Mahmud , J Chen , M Zhou , J Nichols, *"Who Will Retweet This? Detecting Strangers from Twitter to Retweet Information"* ACM Transactions on Intelligent Systems and Technology (TIST), v.6 n.3, May 2015

[13] Bruns, Axel. "Ad Hoc innovation by users of social networks: The case of Twitter." ZSI Discussion Paper 16.2012 (2012): 1-13.

[14] Bruns, Axel, and Jean E. Burgess. *"The use of Twitter hashtags in the formation of ad hoc publics."* European Consortium for Political Research (ECPR) General Conference 2011

[15] Chatfield, Chris. The analysis of time series: an introduction. CRC press, 2016, chapter 2, section 3.

[16] Rozenfeld, Hernán D., et al. "Laws of population growth." *Proceedings of the National Academy of Sciences* 105.48 (2008)

[17] Gabaix, Xavier, et al. "A theory of power-law distributions in financial market fluctuations." *Nature* 423.6937 (2003): 267-270.

[18] Luscombe, Nicholas, et al. "The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties." Genome biology 3.8 (2002)

[19] Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law." *Contemporary physics* 46.5 (2005)