Understanding User Communities from Social Network Data

Alastair J. Gill
Department of Digital Humanities
King's College London
London, UK
alastair.gill@kcl.ac.uk

Emma Tonkin
Department of Computer Science
University of Bristol
Bristol, UK
E.L.Tonkin@bristol.ac.uk

ABSTRACT

Understanding user communities and identifying change within them is important for a range of organisations, including those concerned with cultural heritage. In this paper we present an exploratory study which uses dynamic social network analysis of posts from the Tumblr blogging site relating to the Tate galleries to observe user community change. In addition, we apply two versions of topic modeling to the text of the posts in order to examine user community concerns and changes within these over time. In general, the most noticeable changes in topics within the user communities tends to occur when there has been a major physical change in the social network, such as an increase in membership, with these new members bringing new concerns and interests. After summarising the findings of our approach in detail, we propose practical methods which could be incorporated in to real time monitoring of user community change by cultural heritage organisations.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g., HCI): User Interfaces (D.2.2, H.1.2, I.3.6); I.2.4 Artificial Intelligence: Knowledge Representation Formalisms and Methods (F.4.1)

Author Keywords

Semantic change, social network analysis, topic modeling, dynamic network analysis

1. INTRODUCTION

Maintaining the accessibility of resources is a key focus for cultural heritage organisations; museums, libraries and archives in particular are increasingly working with digital objects, and this has resulted in a growing literature on digital preservation [10]. However, while the challenges of the various formats, software, and hardware changes are well recognized, we are also beginning to understand the impact of semantic change in accessing resources [5]. In addition, a closely related issue is that of different user communities accessing the same resources or the risk of a user communities changing over time. This may present an even greater problem for preservation, since different communities are likely to wish to access information or objects in slightly different ways, for example using different search terms (specific to the different communities), or for a slightly different purpose, perhaps to look for ideas, inspiration and enjoyment rather than simply to identify the provenance of objects.

In this paper we investigate the use of online social networks as a way to begin to empirically study and understand user communities around a cultural heritage organisation. Our motivation is the elicitation of information that enables us to identify and support these user groups effectively at present and into the future. We do not view 'user community' as a single or static entity, and therefore expect that such methods are well suited to understanding such complex phenomena, as well as providing a tractable monitoring method that could be implemented in future.

Social networks may be observed over time to observe and analyse changes occurring within that network [14] . Such an approach has been adopted for a variety of purposes, such as observation of social structures within a given social or organisational context (ibid.), observation of change in attitudes and study of the factors influencing society [9]. A further topic of interest is the spread ('diffusion') of information within social networks [3] and the use of network analysis to explore social influence within a community [19]. The structure of the network itself is a focus of these studies.

In some instances, a detailed analysis of the material contained within the social network is also beneficial. Snapshots of content may usefully be categorized via a topic model [11]. This also permits comparison of the coverage of different sources. For example, Zhao [21] et al make use of unsupervised topic modeling to characterize material within Twitter, then compare with a traditional news corpus, permitting an empirical comparison of the material stored within each. This method allows for comparison between the two corpora. Topic modeling has also been used with to overlay social network data with combined topic and location information [16].

The evolution of topics or coverage over time is less commonly studied. An extended topic model such as the dynamic topic model [7] may be used to analyse the evolution of topics over time in a document collection (e.g., the topics covered over 120 years of publication of *Science*; [7]). An associated approach using a nonnegative matrix factorization framework is described by Saha et al [18] for use on social media content derived from a single platform and on a traditional media source. The use of these methods on data sourced from multiple social media platforms is not widely reported.

Here, we adapt these methods, with a particular focus on the comparative method described by Zhao [21], in order to explore online communities with an interest in a particular art institution, specifically the Tate galleries in London. Being able to identify change in this community is important to preservation for

assessing the social and cultural context of risk, in particular, it is important for the institution (as well as larger cultural and government agencies) to be able to monitor and manage who their audience is for access to the institution and its resources [17]. We use social media for the monitoring of social context with a view to mitigating risk resulting from changes in social context. In the case of the Tate user community identified using social media data, this is largely self-selecting, and we therefore expect it to be fluid and dynamic; any changes are likely to evolve over time.

Based on this analysis, we can identify two primary forms of change: (1) the growth (or contraction) of the community (i.e., the properties of the social network); (2) the change in the concerns/interests (i.e., behaviour) of the community.

To analyse the structure of the social media user community around Tate, we harvest data from Tumblr. In contrast to data derived from the very commonly used micro-blogging service Twitter, access to Tumblr data is not directly temporally limited, and so all historical material remains accessible (with the exception of content deleted or removed by the authors, and of deleted user accounts). This platform is therefore better suited to investigating potential user community changes over time [12]. In the following, we first describe the Tumblr data collection process and resulting data; we then describe the social network analysis and properties of the Tate Tumblr network; next, we describe our analysis of the Tumblr posts from the Tate network. After summarising and discussing these results together, we then propose potentially useful measures for incorporating into any form of monitoring community change within the Tumblr data relating to Tate.

2. DATA COLLECTION METHOD

2.1 Data collection and processing

Tumblr posts were collected for the study of social media content, using the search interface (authenticated via OAuth), which returns query responses in JSON (easily interpretable through compatible libraries such as Python's simplejson). Our general strategy for retrieving relevant content from Tumblr was to initially maximise recall (the fraction of relevant instances retrieved, see [13]) over precision (i.e., the relevance of the posts, which was handled by later filtering). As noted previously, such a large data set of unstructured text/HTML in addition to potentially irrelevant items means that the processing and cleaning up process is non trivial, necessarily iterative, and thus not optimised [1].

For the Tumblr data, a search was completed for posts of any age containing the term 'Tate'. Of the original 70,000 posts from 01-Feb-2005 to 24-03-2015, 3,093 were examined in this present process. This implies that the snowball sampling method applied on Tumblr provides a far lower rate of recall than the keyword search strategies applied on the other two sites; this is unsurprising since snowball sampling depends on serendipity and tends to give a low rate of precision in a highly connected and diffuse network: most of the blogs identified on Tumblr are not solely about Tate, or even art in general, and many of them also link to material that has no relation to either.

The term 'tate' and expression 'tate gallery' were used for the initial search; due to Tumblr's design, it is convenient to search for specific tags rather than solely for terms, and therefore a snowball methodology is used to spiral outwards from initial hits to other posts or blogs that may be of relevance [2,4]. Use of the substring 'tate' in case-insensitive search system captures tagged posts, as well as mentions of the term itself (although we note some false positives were retrieved).

For the purposes of the present evaluation we apply strict filtering rules in order to limit material returned to material containing either the string 'tate' with appropriate word boundaries, or material containing the Tate's hostname. For the Tumblr data, a search was completed for posts of any age containing the term 'Tate'. Of the original 70,000 posts from 01-Feb-2005 to 24-03-2015, 3,093 were examined in this present process, textual data from the 'body', 'caption', 'description' fields gave a total of 473,680 words, 2,793,500 characters (excluding HTML and with URLs normalized; for the topic modeling analysis, further processing [stopword removal and stemming; using the 422 function words from [15] and the Porter stemming algorithm (http://tartarus.org/martin/PorterStemmer/), respectively] resulted in a corpus of 286,360 words; 1,838,736 characters). With contrast to the two other social media platforms and search term approach used ('tate'), we note that the snowball sample method used gives lower precision than the approaches used for the other social media platforms (e.g. for Twitter, 22,00/222,356, or ~10%); however, this is unsurprising given the highly connected and diffuse characteristics of these social media networks, in many cases Tumblr posts linked to material not solely about Tate nor relevant to art in general.

2.2 Describing Tumblr data

In order to give a sense of the Tumblr data, here we present an overview of the qualitative analysis of a random selection of 10 posts, taken from the 3,093 identified as relevant..

Roman Ondak's Measuring the Universe at Tate St Ives stem from the idea of parents measuring and marking the height of their children on the door frame. Love the concept. Art with a temporary existence.	Mountain Lake demonstrates Dalí's use of the multiple image: the lake can simultaneously be seen as a fish. By such doubling he sought to challenge rationality. The painting combines personal and public references. His parents visited this lake after the death of their first child, also called Salvador. Dalí seems to have been haunted by the death of his namesake brother whom he never knew. The disconnected telephone brings the image into the present by alluding to negotiations between Neville Chamberlain, the British Prime Minister, and Hitler over the German annexation of the Sudetenland in September 1938.
A	B

Figure 1. Tumblr excerpts.

These were as follows: four posts were images of art objects (one was a personal image with text containing personal opinion, as in e.g., Figure 1: A; the other three are images and descriptions copied directly from the Tate's collection resources [e.g., Figure 1: B]); in addition, two posts relate to publicity (press releases, interviews, news items on Tate's site) published by Tate relating to exhibitions, and another two further posts mention Tate with reference to broader discussions about the art world; one additional post relates to a personal story of a visit to the Tate, and the final post is irrelevant (it mentions someone named 'Tate'). Of the relevant posts, these results can be summarised as follows: Four posts relate directly to art objects, with five more generally relating to the Tate organisation

3. NETWORK ANALYSIS

3.1 Network analysis method

Network analysis was performed on the cumulative Tumblr data at 6 monthly intervals. Links (edges) between users relevant to the Tate community were identified on the basis of Tumblr connections, with data processed using custom software created in Python. Note that these are shown only from 2009, since before this date there were not enough suitable connections between nodes. Network statistics were calculated at the relevant intervals using the R software package and were average betweenness, number of clusters, density, and average degree. These are shown in Figure 2 (to account for the variation in values across the statistical measures, the y axis is shown as a logarithmic scale; raw counts of clusters can be found in Figure 3). Network graphs describing the Tumblr data are also shown in the illustration above (Figure 1b), with clusters labelled with node user names where relevant. Note that edges are here illustrated as grey lines, and that in cases where there is a high density of connection between nodes, this shows up as areas of solid grey (e.g., 7/2014).

3.2 Network analysis results

Examining the networks of Figure 1b first, we can see that after a very small start to the network in 1/2009, this quickly grows with a greater number of users later in that year, however it is not until mid 2010, that the nodes increase and in turn form more clusters. This pattern continues in 2011, which can be seen more obviously in terms of increase in number of clusters in Figure 3).

The year 2012 however sees a massive growth in the number of users (nodes) in the Tate Tumblr network, and with this an increase in the number of clusters; with this pattern clearly captured by the average betweenness measure showing a great increase in the connectedness of the nodes, and its effective inverse, network density which shows a reduction (Figure 2). From 2012 until the end of our time period (mid 2014), network grown continues, but at a much more steady pace (although note that this is illustrated using a logarithmic scale), reflected in average betweenness and network density; in general the remaining network measure, average degree, is shows a steady increase throughout the whole time period except for two bumps relating to network growth in 2009 and 2012 (Figure 2). Considering Tumblr cluster counts in isolation (Figure 3): here we can see that clusters in general relate to an increase in network size (e.g., 2012, 2014), however this measure is noticeably erratic (e.g., the trough in mid 2012 after the peak in early 2012).

Viewing these measures as a whole, we can see the development of a network relating to a Tumblr community sharing an interest in Tate. Although to a lesser or greater extent, average betweenness, density, and average degree are a function of network size (nodes and edges), we note that there is some utility in the number of clusters within the network for better

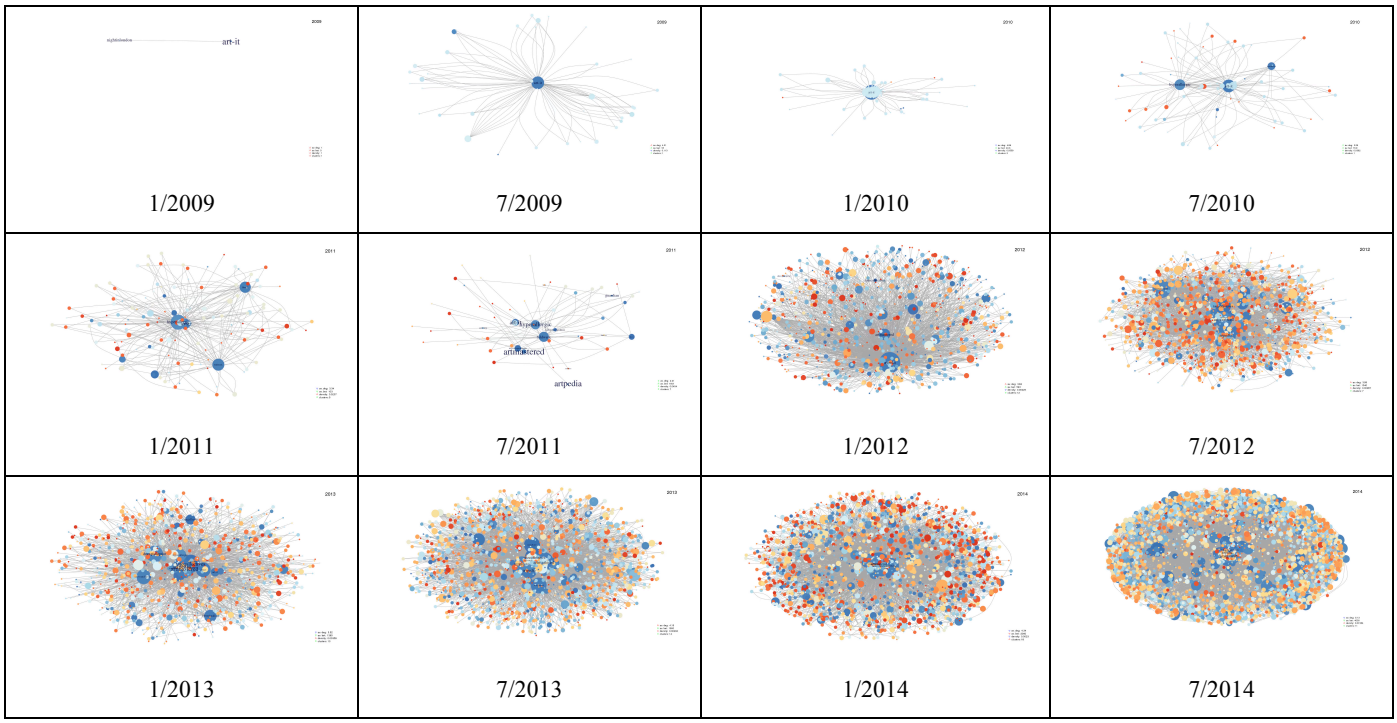

Figure 1b. Tumblr network growth over time.

understanding its structure. However, as the erratic pattern of Figure 3 attests, in practice care may need to be taken in interpreting changes in this measure. In the next section, we supplement the view of network change identified statistically, with analysis relating to the content of the Tumblr community, and how this can help us to understand change.

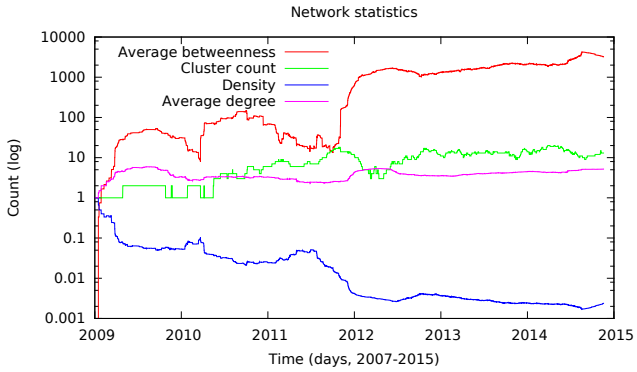

Figure 2. Tumblr network statistics.

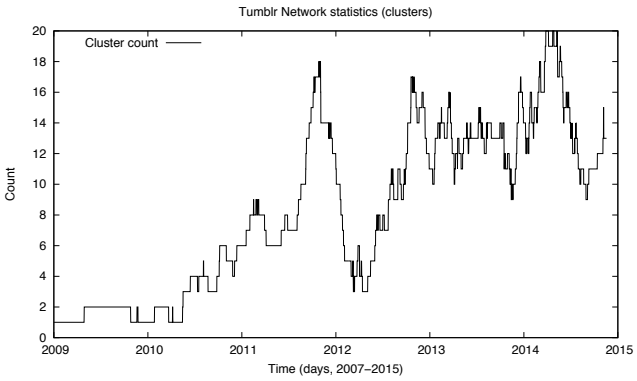

Figure 3. Tumblr network cluster count.

4. TOPIC MODELLING

4.1 Topic modelling method

Analysis of content can provide better understanding of communities in the Tumblr data. Here we adopt an unsupervised machine learning method – topic modelling – which we apply to the text of Tumblr posts in order to identify broad themes. Topic models aim to uncover hidden thematic structures or 'topics' that occur in a collection of documents utilising unsupervised machine-learning techniques [6]. A topic consists of a cluster of words or phrases that show similar patterns of occurrence; documents may relate to more than one topic, and topic modelling calculates a weight with which each topic relates to a particular document. We used Latent Dirichlet allocation (LDA) for topic modelling [6]. As a generative technique, LDA starts with a model that is then used to describe the data by adjusting the parameters to fit the model. The assumption is that the whole corpus of documents contains k number of topics (specified by the user), and that each document talks about these k topics (to a greater or lesser extent). Therefore, each word in a document depends on both the topics selected for that document as well as the word distribution within each of these topics. This intuition is operationalized as a Bayesian Network that models this document generation process.

Since topic modelling is a probabilistic method there are several possible solutions to representing the data, with this process requiring input from the researcher in an iterative process. Here we briefly describe our method: Using the processed data from 'body', 'caption', 'description' fields as described above [cf. 20], topic modelling was performed on the 3,093 documents using the LDA package in Mallet (http://mallet.cs.umass.edu) to generate models for a variety of number of topics, ranging from 3-20 (3, 5, 7, 10, 15, 20). This was in order to select a number of topics which best describe the data. In all topic modelling described here, default parameter settings were used except in the case of α where a relatively low value (0.01) was specified in order to generate topics which relate more distinctly to particular documents, cf. [8]. Two researchers familiar with the data set visually inspected the model outputs to evaluate the 'topic keys' (words most representative of the topic), to determine whether they contained a disproportionate number of poor topics which would indicate a poor description of the data (specifically topics which were too general, too specific, repetition with other topics, or internally inconsistent). These key words were then used to manually search through documents containing the respective topics in models considered suitable for our analysis, in order to get a better sense of that topic for interpretation and naming of the topics. Following this process we settled on two topic models which appeared to provide a good summary of the data: these specified 5 and 15 topics, and they are shown (with their topic keys) in Tables 1 and 2 respectively. In the following section, we describe these in more detail in relation to the data.

4.2 Topic model descriptions of Tumblr data

4.2.1 5 Topic solution

The five topic description of the Tumblr data shows four contentful and more frequently used topics, with the fifth (Foreign) showing much lower usage (and which relates mainly to non English texts): The two topics with a similarly great level of use are URL/Modern and URL/ArtworkProperties (0.32 and 0.31, respectively). Of these, the first relates mainly to the Tate Modern (e.g., exhibitions, or passing references to the gallery), with the second relating to the physical properties of artworks (e.g., factual information such as canvas size or materials used). Both of these topics feature URLs which presumably are being referenced by the author and are – to a greater or lesser degree – the subject of their post.

The topic with the third greatest proportion is IMG/description (0.26), which contrasts nicely with the previous topic URL/ArtworkProperties, as it refers to a linked image (which the post relates to), and a description of the art object, however rather than being concerned with the factual physical and material aspects, the IMG/description topic gives a personal perspective and interpretation of the art object (e.g., the flow of the brushstrokes, or the meaning of the scene). The final topic, ExhibitionInfo (0.20), provides information about exhibitions, perhaps especially promotional material advertising exhibitions.

4.2.2 15 Topic solution

The 15 topic description of the data can be grouped into four main clusters based on the proportion of topic usage in the Tumblr data, although as may be expect given the greater number of topics, the proportions found for topic usage are lower than for those of the 5 topic solution: The first group with proportions of 0.20 or greater

within the data are IMG/PaintCharacteristics (0.22), IMG/Exhibition/Art (0.21), URL/Description/Materials (0.20), and URL/Modern/Performance (0.20). The first topic mentions descriptions of paint, techniques and characteristics in relation to an image(s), the second again contains an image as well discussing exhibitions in relation to art, the third gives a factual description of an art object in terms of materials and referencing a URL, and the fourth most frequently used topic includes a URL along with content relating to (Tate) modern and performance.

Table 1. Five topic model of Tumblr data

Topic Label	Topic ID	Key Items	Proportion
URL/Modern	4	removedurl tate modern art london week museum exhibit matiss	0.32
URL/Artwork Properties	0	removedurl paint tate work paper cm removedimg canva artist	0.31
IMG/description	1	removedimg work exhibit show tate art piec time paint	0.26
ExhibitionInfo	2	art work artist exhibit tate museum modern galleri perform	0.20
Foreign	3	de video art la le artist pari film en	0.02

The second more frequent grouping of topics (with proportions of between 0.10 and 0.19), are Exhibition/Film (0.14), Descriptions/Britain/URL (0.14), 3D (0.10). The first of these topics contains reference to exhibition along with mention of film, the second topic contains factual catalogue-type information relating to (Tate) Britain along with a URL, and the third topic of this group contains materials and characteristics of three dimensional art objects.

The third group have usage proportions of between 0.5 and 0.9 in the Tumblr data, and are ArtworkContext/URL (0.08), Modern/Artists/URL (0.08), URL/StIves/Landscape (0.06), URL/Exhibition/London (0.05), and URL/Modern/Podcast (0.05). Of these topics, the first describes the context of the art object more generally (and less prominently a URL), the second relates to (Tate) Modern and names of artists from the past 20 or so years (and also a URL), the third to some extent mentions (Tate) St Ives, as well as other concepts such as 'landscape', and the final topic contains a URL along with (Tate) Modern, which appears to be in the context of exhibitions, and also mentions 'podcast'.

Finally, the three least used topics are Cities (0.02), Video (0.01), and Foreign (0.01), which are respectively, mentions of the word 'city/ies' or names of cities, references to video (as in art object, but also videos posted on social media), and non-English words.

Comparing the two topic models run on the Tumblr data, it is unsurprising that the 15 topic solution provides greater granularity than the 5 topic model. However, what seems more apparent between the two models is that the 5 topic model is more abstract, giving a better sense of the concepts and types of posts it is describing (e.g., description of an image, materials used, advertising an exhibition), whereas the 15 topic model provides more information about specific content (whether content relates to 3D, Tate Modern artists, St Ives, Tate Britain). Both of these models, and their respective granularity, have advantages and disadvantages when describing the data in light of identifying community change. We discuss this briefly, below.

Table 2. Fifteen topic model of Tumblr data

Topic Label	Topic ID	Key Items	Proportion
IMG/Paint Characteristics	10	paint work artist removedurl removedimg imag figur colour form	0.22
IMG/ Exhibition/Art	1	removedimg work show time exhibit make veri year love	0.21
URL/Description/Materials	12	removedurl tate cm paper oil canva sourc collect removedimg	0.20
URL/Modern/ Performance	6	removedurl tate modern removedimg artist music perform sound tank	0.20
Exhibition/ Film	9	art work tate exhibit artist museum film present perform	0.14
Descriptions/ Britain/URL	14	art exhibit galleri london tate removedurl artist britain british	0.14
3D	3	space sculptur piec galleri yellow black build tate rothko	0.10
Artwork Context/URL	0	art cultur artist removedurl work polit commun peopl beui	0.08
Modern/ Artists/URL	7	kusama tate hirst modern yayoi removedurl damien exhibit room	0.08
URL/StIves/ Landscape	5	st iv removedurl landscap war tate sea mso picasso	0.06
URL/ Exhibition/ London	13	removedurl galleri london august novemb matiss tate modern lincoln	0.05
URL/Modern/ Podcast	11	removedurl art modern week podcast museum exhibit matiss barlow	0.05
Cities	4	citi al eliasson removedurl walk york film london project	0.02
Video	2	video art artist instal exhibit paik de pari includ	0.01
Foreign	8	de la le en du dan art au par	0.01

5. EXPLORING TUMBLR CONTENT OVER TIME

So far we have shown how social network analysis provides information about the size and relationships between the Tate community identified on Tumblr, and topic modelling provides information about the content of Tumblr posts. In this section we provide analysis of how the Tumblr topics identified in the previous section change over time, and how these can be related back to the network changes identified in the first section; this will in turn give us a better understanding of user community change, especially in terms of how their concerns – expressed through Tumblr posts – change over time. In addition to describing the results, we will provide examples of community change identified in this data, and in the final section describe how these can be incorporated in to a automatic process to identify community change .

Topic relationships to documents used to create the two models in the previous section were used to provide the primary topic representing each Tumblr post (i.e. the single topic showing the highest proportion of usage in each document); these were then summed for each 6 month period (January-June 2009, July-December 2009, January-June 2010, etc. until the final period in our data collection, January-June 2015). We note that 4 posts used in the previous generation of the topic models were excluded since they dated from before 2009; this left 3,089 remaining posts (as per the social network analysis, above). Usage of content described using the 5 topic model is shown in Figure A4, and the 15 topic model is shown in Figure A5; the y axis is the frequency of posts which are primarily described by that topic.

Both the 5 and 15 topic models of the Tumblr data over time show the rapid increase in posting activity in 2012, reaching a peak in the latter half of that year; activity remains high over the next two years, albeit generally declining; the first half of 2015 has relatively low Tumblr posting activity for the Tate community, but this may be due to an incomplete collection of data for this period. The peak of activity identified in 2012 appears related to the massive network growth also shown in the social network metrics of Figure A2. Using this data as an example, we now explore how this community growth in 2012 and after, is reflected in the change of content in the Tate Tumblr community.

Looking in particular at the Tate Tumblr community use of the 5 topics over time, we note content usage changes as follows: From 2012 onwards, we see a decline in use of URL_ArtworkProperties (relating to catalogue descriptions of art objects), which picks up again in 2014; with this dip in usage coinciding with a bump in IMG_description (which relates more to images and their exhibition rather than catalogue data) in late 2012 and 2013. In addition, there is a consistently higher use of ExhibitionInfo following 2012 (giving details of exhibitions without relation to a specific image), and a relatively higher proportion of posts primarily concerned with URL_Modern (perhaps indicated a higher profile of the Tate Modern and modern works within the collection). We note the very rare use of primarily Foreign content, with the exception of the latter half of 2010.

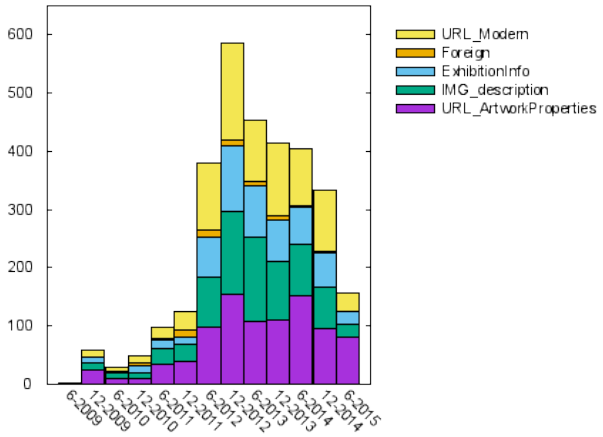

Figure 4. Tumblr topic frequency over time (2009-2015) (5 topic model).

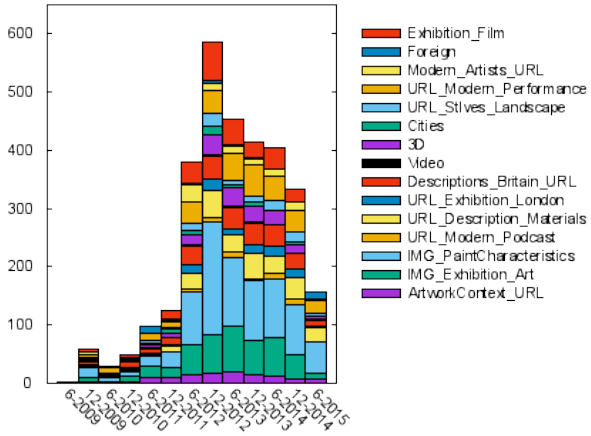

Figure 5. Tumblr topic frequency over time (2009-2015) (15 topic model).

In summary, the 5 topic model shows that this network and resulting community change appears to indicate an apparent initial focus on images relating to exhibits, but this appears to have been used in place of posts sharing catalogue information of art objects. This change has also resulted in greater sharing of exhibitions at the Tate in reference (and promotion?) to the Tate Modern.

In terms of the 15 topic model, we find that posts primarily about IMG_Exhibition_Art and URL_Modern_Performance dip around 2012, with Modern_Artistists_URL, in contrast, peaking around this time; in general, IMG_Paint_Characteristics and URL_Description_Materials posts increase in proportion and stay relatively more common from 2012. We note that the remaining topics were used relatively infrequently or inconsistently across this time period; we do not discuss them in detail.

In summary, the 15 topic data appears to reveal that the rapid network increase around 2012 temporarily focused on Modern_Artists_URL (posts of links relating to Tate Modern artists), while posts relating to exhibitions in general

(IMG_Exhibition_Art) and performances at the Tate Modern (URL_Modern_Performance) descreased around this time; this network and resulting community change resulted in a continuing, greater number of detailed posts focussing on the physical properties and painterly aspects of art objects (IMG_Paint_Characteristics and URL_Description_Materials), which in practice may be the popular describing and critiquing of objects in the Tate catalogue. More generally, we see that the granularity of the 15 topic model provides detailed topics which come in and out of usage, rather than the fairly consistent usage (albeit with variations in proportion) found in the 5 topic model.

Therefore by combining social network analysis metrics to the Tumblr network and topic modelling to the content of the posts of the Tate community on Tumblr, we have identified an example of change in this community relating to the growth in and around 2012: We note that both the 5 and 15 topic models identified this change in the content generated by the Tumblr community in relation to Tate; in particular, the adaptation of this social network and its content to meet its new needs. The 5 topic model identified a temporary change in focus from catalogue data to image data, and a greater focus on the Tate Modern and sharing exhibition information. Although the first two topic changes may indicate an exploration with new media, it is the focus on Tate Modern by the community and sharing/promotion of exhibitions which seems to indicate a more substantive shift in community usage of Tumblr.

For the 15 topic model, although many of the topics are used infrequently and which come and go in usage, in this analysis example we focused on five. From this example analysis, we found that following 2012 there was an increase in the popular describing and critiquing of art objects, along with a temporary focus on Tate Modern artists, and similarly less focus on images relating to exhibitions and performances at Tate Modern. Of these, we note that the change of focus relating to Tate Modern artists rather than exhibitions is interesting, and provides more detail to the general increase in posts relating to Tate Modern identified in the 5 topic model; in contrast, the increase in description and critique of art objects captured by the 15 topic model is only regarded as a temporary change in exploring the use of image descriptions in the 5 topic model. Regardless of these nuances, we view these broad changes as the increase in number of art appreciation posts, as well as an increased interest in the community relating to Tate Modern. Both of these large scale changes of community behaviour are indicative of a social and cultural context, which we expect to be important in understanding the Tate in its broader online and offline community context.

Overall, the results from the two models show similar changes in the Tate Tumblr community (primarily the description of art objects and coverage of Tate Modern), but their different granularity and probabilistic generation mean that they provide detail in different ways, in some cases identifying increase of a topic, and in others the change in use from one topic to a similar one, but with nuanced differences. This would indicate therefore that at least for initial monitoring purposes, it would makes sense to include the topics from both models in this process, thereby allowing the greatest insight into community change processes; the disadvantage to this is that there would be a slightly greater amount of data to consider, but in this case it does not seem to be too arduous, given that this would result in 20 topics in total. As noted before in relation to the 15 topic model, some of these topics occur with a relatively low frequency in the Tumblr data – this may lead to the possibility that such a model over fits the data, however, given that we propose the inclusion of the 5 topic model, then we expect this risk to be mitigated by the use of the broader topics, and greater coverage that this smaller model provides.

6. APPLICATION AND DISCUSSION

In the following section we now discuss ways in which the approaches of this paper could be integrated into future community change risk assessment. These are covered in turn:

Changes in network properties: these are important to identify, since as found in the example reported above, the growth of a network may result in different community uses of the social network (i.e. sharing different types of content for different purposes). The network metric data presented in Figures 2 and 3 is per day, and so shows a great deal of variability; smoothing data – for example by calculating average metrics over the previous 6 months – would provide a more stable measure of social network metric change. In particular, average betweenness (connections between nodes) and number of clusters (groupings within the network) appear to be a feasible way of identifying social network change in this data set. Using smoothed data, it would be reasonable to monitor social network change based on an average of the previous 6 month's data in comparison with e.g., greater than 1 standard deviation difference from the mean of the previous 2 year's data. This metric could be tuned in order to ensure that the great network change which occurred in 2012 is identified, but that more random change such as network volatility found from 2009-2011 is largely ignored. Testing and further analysis would be required to explore whether the gradual and steady growth in the network between 2013-2015 should be identified as change in this model.

Changes in post content: although network change happens, it is necessary to identify how this change has an impact upon the community. Here we use probabilistic topic models generated on the data to classify the content of the Tumblr post into broad categories. As our discussion of the results (Figures 4 and 5) shows, not all of the topics are important for, or related to, an instance of community change. However, since we do not know what topics are relevant in advance, it is important that the results of all topics is available for review when network change is identified. Because of this issue and the interpretation required in understanding community change based on post content, we note that this process should be performed by a human with expertise in this domain. This review process can be automatically assisted, with the topics most relevant to community change highlighted. We anticipate that similarly to the identification of network change, that this should be based on current topic proportion (for the previous 6 months) compared with the 2 years previous (with difference highlighted in descending order of standard deviations difference between the current and previous time points).

7. CONCLUSION

In this paper we have investigated an important aspect of change in relation to digital preservation, namely community change. This is important since the social and cultural context in which a cultural institution operates, determines how it can serve its community. In particular, we have addressed this question using social network data harvested from Tumblr relating to Tate over the period 2009-2015.

In addition to exploring changes in social network relationships over this time, we have built a probabilistic model of the textual content of Tumblr posts using topic modelling. Comparing two models identifying different granularity of content over this 5 year

period, we have been able to identify community changes in the use of Tumblr content in relation to an example of major network growth in 2012. This network growth resulted in different focuses by the Tumblr community, with these mainly related to long standing changed of an increase in the proportion of posts relating to Tate Modern and the popular critiquing and presenting to the community of Tate art objects; the different topic models each provided differently nuanced perspectives on these behaviours.

Based on the findings of this example analysis, we propose how this could be incorporated into the automatic monitoring of community change for risk assessment. Here we detail two metrics and thresholds which could be used for community change monitoring, namely changes in network properties and changes in post content: we anticipate that the former will be used to identify large scale community changes using proposed thresholds based on the current change example analysis; then the relative changes in topic usage over time will be presented for assessment by a domain expert, since this evaluation will necessarily require human interpretation and knowledge. We also expect that human evaluation will be required in order to ensure that appropriate network change thresholds are being used.

In relation to the example analysis of the Tate community Tumblr data presented above, we noted that the insights provided by such analysis, for example the growth in the network and the resulting increase in posts relating to art appreciation posts, as well as an increased interest relating to Tate Modern, both provide important indications of the wider social and cultural context of Tate. Such information is instrumental to development of an improved understanding of the broader online and offline community context in which the organisation's activities are received.

8. ACKNOWLEDGMENTS

The authors acknowledge funding from the PERICLES project via the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 601138.

9. REFERENCES

[1] Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012, June). Semantics+ filtering+ search= twitcident. exploring information in social web streams. In Proceedings of the 23rd ACM conference on Hypertext and social media (pp. 285-294). ACM.

[2] Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. Social research update, 33(1), 1-4.

[3] Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web (pp. 519-528). ACM.

[4] Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. Sociological methods & research, 10(2), 141-163.

[5] Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. Historical semantics and cognition, 13, 61-89.

[6] Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM

[7] Blei, David. (2012). Topic modeling and digital humanities. Journal of Digital Humanities 2 (1),8-11.

[8] Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. Handbook of Mixed Membership Models and Their Applications, 3-34.

[9] Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. New England journal of medicine, 358(21), 2249-2258.

[10] Hedstrom, M.. (1997). Digital preservation: a time bomb for digital libraries. Computers and the Humanities 31, 3 (1997), 189–202.

[11] Hu, Y., John, A., Wang, F., & Kambhampati, S. (2012, July). ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback. In AAAI (Vol. 12, pp. 59-65)

[12] Jain, P., Kumaraguru, P., & Joshi, A. (2015, August). Other Times, Other Values: Leveraging Attribute History to Link User Profiles across Online Social Networks. In Proceedings of the 26th ACM Conference on Hypertext & Social Media (pp. 247-255). ACM.

[13] Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. Information storage and retrieval, 7(5), 217-240.

[14] Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. science, 311(5757), 88-90.

[15] McDonald, S.(2000). Environmental Determinants of Lexical Processing Effort – Unpublished Ph.D. thesis, University of Edinburgh.

[16] Pozdnoukhov, A., & Kaiser, C. (2011, November). Space-time dynamics of topics in streaming text. In Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks (pp. 1-8). ACM.

[17] Schlieder, C. (2010). Digital heritage: Semantic challenges of long-term preservation. Semantic Web, 1(1, 2), 143-147.

[18] Saha, A., & Sindhwani, V. (2012, February). Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 693-702). ACM.

[19] Tang, J., Sun, J., Wang, C., & Yang, Z. (2009, June). Social influence analysis in large-scale networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 807-816). ACM.

[20] Wang, Y-C, Burke, M. and Kraut, R. (2013) Gender, topic, and audience response: an analysis of user -generated content on Facebook. In Mackay, W.E., Brewster, S., and Bødker, S. (Eds.).Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13) (pp. 31-34) New York, NY: ACM.

[21] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In European Conference on Information Retrieval (pp. 338-349). Springer Berlin Heidelberg.