Tweet patterns reveal community membership: a step towards measuring integration on Twitter

Lewis Lloyd

with thanks to Luc Berthouze.

Abstract:
Measures of integration in society have yet to take into account how people behave and interact online, and particularly on social media such as Twitter. One of the main difficulties involved in developing such a measure would be the effective identification of sub-communities within a population. Starting with a population of tweeters, we demonstrate the possibility of disentangling the contributions of different sub-communities to the behavioural profile of that population. More specifically, we develop the beginnings of a methodological framework for an approach to grouping users based solely on correlations in their tweeting patterns over time. This idea, while simple, is seemingly novel. We demonstrate both its efficacy as a standalone method for identifying communities, and encourage its use in combination with the more common content- and interaction-based methods for identifying related Twitter users. Finally, we consider ways in which our approach could be used to develop more quantitative measures of online integration.

Contents:

1. Introduction

Measures of social integration abound. The extent to which diverse populations live in the same areas, send their children to the same schools, share the same attitudes and socialise together is often assessed (Casey, 2016; Heath et al., 2013; DCLG, 2012). However, little attention has been paid to the virtual world where people spend an increasing proportion of their time, and the social media on which they conduct an ever-larger part of their social interactions (Asano, 2017). For assessments of integration to accurately reflect the lives people lead in the 21st Century, this needs to be addressed.

Twitter is one of the dominant platforms in this virtual world. An average of 328 million monthly active users (Statista, 2017) generate roughly 6000 short statuses ("tweets") per second, resulting in over 500 million per day (Twitter, 2013). As well as sending their own tweets, users can follow the activity of others, "favouriting" and "retweeting" tweets that they like. While some opt to keep their accounts private and viewable only to followers, the majority of this content is in the public domain and accessible through application programming interfaces (APIs) provided by Twitter. As a result, considering behaviour on Twitter would seem a sensible starting point for developing measures of virtual integration, and it is accordingly Twitter that we focus on here.

Attempting to measure social integration assumes two things: that you have a population to work with, and that communities of interest within that population have been identified. A population might then be deemed integrated by virtue of the existence of strong social ties or similar patterns of behaviour between sub-groups. Equally, the integration of a community into the whole could be assessed on the basis of the similarity of its activity to that of other groups, and how far its members interact with people outside of said community.

While we address the question of identifying a suitable population to work with in Section 3.1, the majority of this study concerns the harder task of identifying distinct communities within a population. Rather than analysing the content of users' tweets or relying on explicit social ties, we wondered whether the interests and outlook of users could be inferred from the events they responded to, with community membership derived from there. Ultimately, patterns of behaviour over time were used as a proxy for event responses, with the reasons underlying this evolution explained in Section 4. The idea that a community might be defined by a shared outlook on the world is not new, building on Émile Durkheim's concept of a *conscience collective* (Durkheim, 2013). To the best of our knowledge, however, our approach to clustering Twitter users based solely on temporal behaviour patterns *is*.

After an evaluation of our attempts to identify communities with a "collective consciousness" in Section 5, Section 6 sees us return to the question of integration. The scope for using our temporal method to gauge levels of integration within populations on Twitter is discussed, with possible approaches to developing a more quantitative measure of integration proposed.

2. Background

Social integration, while difficult to define, can be seen broadly as the extent to which the members of a society share common spaces and values. Long a subject of study in the social sciences, measures have yet to evolve to take into account behaviour in virtual space. Research has considered the extent to which diverse communities use online services differently (Khoo, 2014), and concern about "filter bubbles" and "echo chambers" in recent years has prompted a range of work assessing how far like-minded users group together on social media (Krasodomski-Jones, 2016; Himelboim et al., 2013; McPherson et al., 2001). However, no prominent attempt has been made to measure the integration of a population on Twitter, nor to link an assessment of virtual behaviour to existing measures of integration.

Such a measure would require the accurate detection of groups of related users within a population. Grouping similar objects is, of course, a very common problem across a variety of domains, and there is a wealth of literature outlining countless different approaches to solving it. Most influential here was work on community detection in social networks, and Blondel et al.'s (2008) detailing of the "Louvain Method" in particular. Not only do we use their method in this work (see Section 4.2.b), but there are clear parallels between their application of community detection on a Belgian phone network to an assessment of the integration of the Belgian population and our project.

In Twitter analysis, too, the problem of identifying communities of users is not new, with most work adopting either one or a combination of two broad approaches (Cihon and Yasseri, 2016). The first places an emphasis on the content of users' tweets and profiles. Keywords, hashtags, sentiment, links shared and more are processed, allowing for the inference of user characteristics such as gender (Burger et al., 2011) income level (Preotiuc-Pietro et al., 2015), age and political orientation (Rao et al., 2010). The second approach is more heavily focused on relations – who follows or is followed by whom – and interactions – retweets, replies, mentions – between users (Tang and Liu, 2010). Clearly, there is scope for both approaches to be informative, and many methodologies blend the two (Varol et al., 2017; Tyshchuk et al., 2014; Bodine-Baron et al., 2016; Himelboim et al., 2013).

The temporal approach outlined here follows a different path, combining research into time series analysis of Twitter (e.g. Bie et al., 2016) with the growing literature on responses to events on social media (Mugan et al., 2013; Akcora et al., 2010). While temporal tweet data is generally overlooked, Bagheri et al. (2015) recognise its potential importance, expanding the typical, topic-based content acquired for users with an assessment of behaviour over time when carrying out community detection. However, as far as we are aware, no previous work has relied solely on temporal data to identify communities of Twitter users.

3. Data collection

3.1 Identifying a population

The starting point for the project was finding a 'population' of Twitter users to work with. One option was to select a population of people in the real world, sharing a physical space or common culture, and find their diverse online presences. Lots of research has gone into the problem of inferring Tweeter location (e.g. Graham et al., 2013), for instance, and this could have been built upon. However, the likely difficulties of accurately identifying such a real-world population risked consuming the project. Moreover, for our exploratory purposes, it seemed unnecessary.

Instead, we opted to identify a population based on shared online spaces, regardless of physical location or cultural or ethnic identification. Margetts et al. (2015) have pointed out that social media enable the development of communities not bounded by geography. These communities might be defined by users interacting as part of the same network, or – given the majority of social media behaviour is browsing (Khoo, 2014) – by virtue of users following similar accounts, and consuming similar content as a result. This would certainly suggest shared values, at least. For the purposes of being able to identify distinct sub-communities, though, it was important that the users in question were not too similar, and we accordingly looked for just one account whose followership could be taken as a population.

The account settled on was the official Leicester City football club account (@LCFC). This was for a range of reasons. Followers of a football club would have at least one shared interest to link them, with the scope for being entirely different on any other metric (given the broad popularity of football). Moreover, it offered a likely physical location for those followers, bringing in an element of more standard definitions of population. That physical location, too, is strikingly diverse in the case of Leicester, with almost half of the city's population not ethnically white, and one third born abroad (ONS, 2012). Furthermore, unlike the Manchester Clubs, Chelsea, Arsenal or Liverpool, Leicester's rapid rise to prominence over the last 18 months from relative obscurity makes it more likely that their followers are based locally, rather than internationally.

3.2 Collecting the data

Given the proposed temporal approach to identifying communities, we were interested in collecting the tweets of each user in the population covering a substantial period of time. Intuitively, the longer that period was, the better, since a prolonged period would offer more scope for correlations between the activity of users to be found. As the Twitter API does not allow for tweet collection covering a given historical time period, collecting as many tweets as possible for each user, then disregarding any outside of a selected period, seemed sensible. However, with access to up to the last 3,200 tweets for each user, collecting for all 1 million Leicester City followers could result in an upper

bound of 3.2 billion tweets. Such an enormous volume of data would be entirely unwieldy, and some pruning of the users to collect for was necessary. For this we used Method52, a system for collecting, filtering and classifying social media data developed by the Text Analytics Group (TAG) Lab and CASM Consulting LLP at the University of Sussex.

Tweeters whose accounts were protected were immediately disregarded, since it would be impossible to access their tweets, along with any for whom the language associated with the account was not English. Given that we were looking for behaviour over time, there was at least a minimum quantity of tweets that would be required to establish any measure of behaviour, and users who had tweeted less than 200 times were accordingly removed. Users who had tweeted more than 10,000 times were also undesirable, on the basis that their last 3,200 tweets may not take us back very far. As a further means of cutting down numbers, while increasing the likelihood of shared location, only users whose profiles were associated with a location in the UK or with Greenwich Mean Time (GMT) as their time zone were kept in the population.

The resulting stream of tweets covered almost ten years. However, the activity of some users went back much further than others. A point was settled on (21st March 2016) to optimise the length of time covered and the number of users whose activity spanned the resulting period (a user qualified if the oldest tweet collected for them was either at or before this point). The 12,000 users identified dropped to 11,982 once any who did not tweet at all in the period were removed. Finally, given that the tweet collector had run for multiple weeks on account of Twitter's API limits, tweets from after the point at which the collection process had started were removed. This left 7,149,346 tweets to work with. The distribution of these over the 14.5 month period between 21st March 2016 and 9th June 2017 is plotted below (Figure 1).

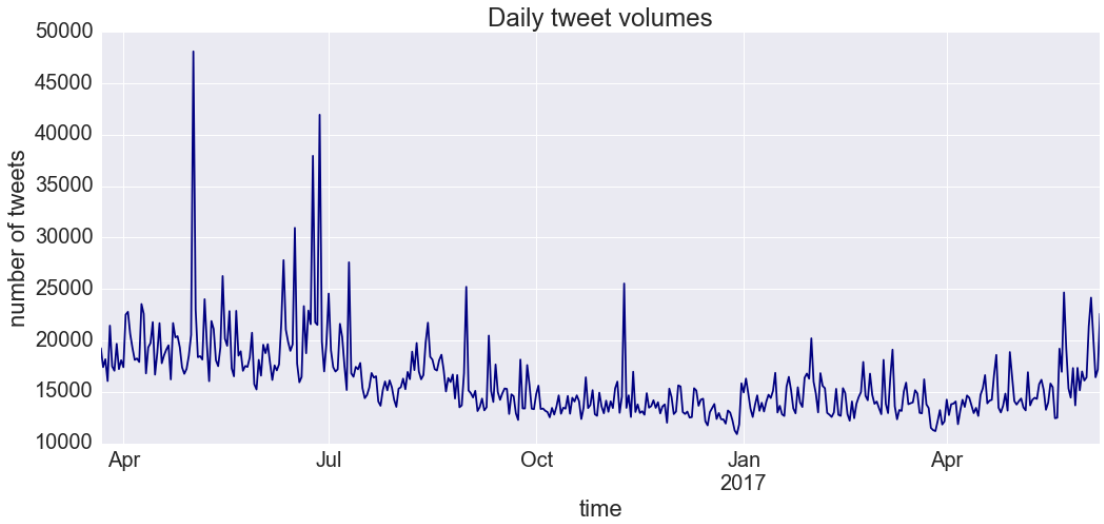


Figure 1: a visualisation of the volume of tweets produced by the population over time. Daily counts are plotted.

4. Methodology

The idea to use temporal data alone came from a prior consideration of event responses. It ought to be possible, we hypothesised, to infer users' interests from the events they responded to (Mugan et al., 2013 have done something similar, for instance). In the case of Leicester City followers, it was expected that most would tweet in some manner about Leicester City football matches. Below this level of common interest, though, it was expected that many would also tweet about the European football championships (Euro 2016), others about political events (during a particularly turbulent time for British politics), others still about musical events (Glastonbury, the Brit awards), and so on. The events could be of any nature, with consistent reactions over time to events of a certain type indicating interest in a certain area. Furthermore, sentiment analysis could be employed to infer an outlook on topics of interest – as Himelboim et al., 2013 do for identifying political ideology – and perhaps even identify points in time at which opinions changed (c.f. Akcora et al., 2010).

However, there were obvious limitations to an approach based on responses to events:

1. The process of identifying the events to focus on would be time-consuming and limiting. Some combination of selecting events based on those the population appeared to be responding to from peaks in activity, and devising a hand-crafted list of events likely to be pertinent could be adopted, but both are labour-intensive and would necessarily (unless every micro-peak in population activity was painstakingly assessed) result in many events being overlooked.
2. Even with events decided upon, establishing the time period during which responses to each individual event should be retrieved poses challenges. Major events often result in a pronounced divergence from baseline activity, and the period can then be taken as the time between the event happening and the point at which behaviour returns to normal. However this point is not always easy to identify. In such cases, should the hour, day, week or month around the event be considered? Should this remain consistent for all events, or for events of the same 'type' – or should it be established individually? With a small corpus of tweets, everything from the event until the end of the period for which data had been collected could be assessed, but for a large corpus this would likely be unfeasible.
3. Developing sufficiently accurate relevancy classifiers for potentially hundreds of events across diverse domains would be even more time-consuming than the above combined – or costly, if you paid someone to do the training for you.

It seemed easier and potentially less limiting to use behaviour over time as a proxy for event responses. If users consistently tweet at the same times as each other, it is likely that they are responding to events of a similar nature. Turning the event-focused approach on its head, if we could identify groups of users with correlated behavioural patterns over time, looking at the peaks in their activity should reveal the sort of events they were responding to, and consequently allow for the inference of community membership.

4.1 Finding correlations between users

4.1.a Representing behaviour over time

To convert our stream of 7 million tweets into a representation of each user's behaviour over time, we used *pandas* – a Python data analysis library (McKinney, 2011). For each user, a time series was created with the moments at which the tweets in our corpus were sent as the index, resulting in a row for each tweet. The values of the time series were set to 1 for moments at which the corresponding tweet had been sent by the given user, and 0 otherwise.

Storing all of these time series would have been impossible (some 85 billion 1's and 0's were involved), and it was unlikely that meaningful correlations could be found between users based on their second-by-second activity even if we could store representations of that behaviour. We therefore used *pandas'* "resample" method to collapse the time series for each user as soon as it was created. The converted index consisted of time intervals instead of moments. The values of the series were calculated as the sum of all of the tweets sent by the user in each interval.

This process was carried out for four different interval lengths spanning the 14.5 month period for which we had data: two hours, one day, three days, and one week. At each level of temporal granularity, the time series for all of the users were concatenated into a *pandas* "DataFrame". By way of illustration, part of the DataFrame representing weekly activity appears below (Table 1), covering the first 6 week-long intervals (rows; the dates shown reflect the end of the intervals), and 6 randomly selected users (columns).

	A	**B**	**C**	**D**	**E**	**F**	**…**
2016-03-27	144	2	10	0	5	58	…
2016-04-03	91	2	12	1	5	48	…
2016-04-10	42	6	5	0	2	67	…
2016-04-17	26	6	8	2	1	55	…
2016-04-24	7	7	10	1	0	110	…
2016-05-01	3	6	15	5	0	30	…
…	…	…	…	…	…	…	…

Table 1: excerpt from *pandas* DataFrame encapsulating users' weekly activity.

The resulting DataFrames stored user activity as continuous values corresponding to the number tweets sent during each interval. However, since we were primarily interested in when people tweeted rather than the amount that they tweeted, an alternative approach was to use categorical values indicating simply whether or not a user had tweeted in a given period. We used scikit-learn's (Pedregosa et al., 2011) "Binarizer" to convert any value greater than 0 in the original DataFrames to 1 to this end, and stored the outputs as separate DataFrames.

4.1.b Measuring similarity

We opted to measure the pairwise correlations between our representations of user activity as the primary means of assessing similarity. As noted above, we were interested in when people tweet, rather than how much they tweet. For our purposes, perfectly similar users should exhibit identical patterns of temporal behaviour: when the activity of one increases or decreases, the same incline of decline in activity ought to be visible for the others – relative to their typical tweeting habits. A measure of how strongly user activity was correlated therefore seemed more appropriate than a measure of distance. The latter would place undue emphasis on the frequency with which users tweet, likely marking users A and F in Table 1 (above) as similar, rather than users A and E, whose tweeting patterns match much more closely.

The resulting distributions of correlation scores are visualised as histograms.

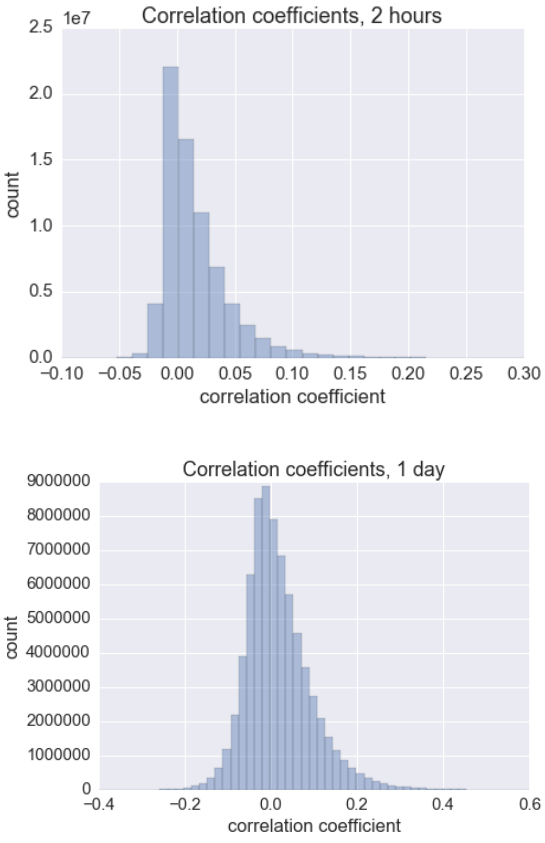


Figure 2: distributions of correlation coefficients corresponding to patterns of two-hourly (top) and daily (bottom) behaviour.

4.1.b(i) Data with continuous values

Correlation coefficients were calculated for the data with continuous tweet values first. Each DataFrame was passed to *pandas'* "corr" function, with the default method – Pearson's, which divides the covariance of the two variables by the product of their standard deviations – used. The function returned correlation matrices, with values between -1 and 1. The distributions of these values are plotted as histograms (see Figures 2 and 4).

Two things are particularly striking from these histograms. The first is that the overwhelming majority of the nearly 150 million correlation coefficients in each case are clustered around 0. This is especially evident for the correlations drawn from the data representing behaviour over shorter time intervals. Only 0.04% of the values from the data covering two-hour behaviour have a magnitude of greater than 0.3 – a threshold below which correlations are typically seen as uninformative. This suggests that working at such a high level of temporal granularity is likely to be unproductive. While the proportion of meaningful correlation coefficients is slightly better for the data on daily behaviour at 0.5%, it remains poor.

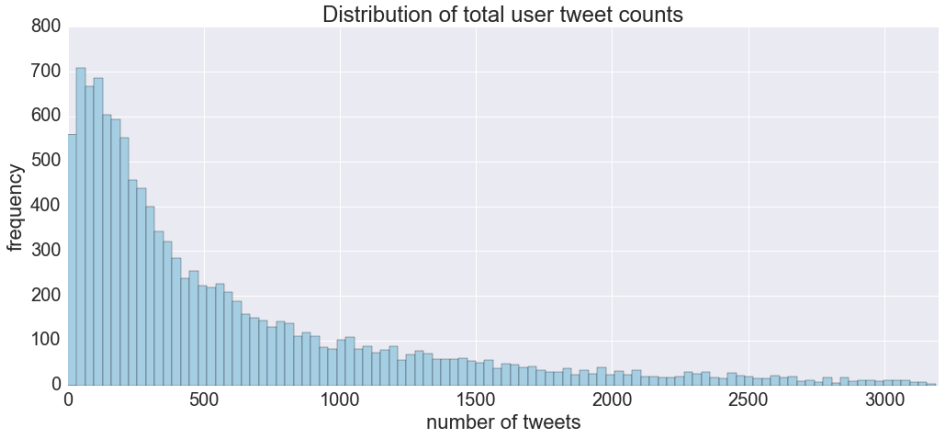


Figure 3: distribution of total tweet count frequencies over the period.

This is understandable when we take into account how often most of the users in our population tweet, which can be deduced from Figure 3 (above). The same trend has been observed for Twitter users in general (Hubspot, 2009), suggesting it is not simply a quirk of our data – and this is even with users with the lowest tweet counts removed (see Section 3.2). Given that the majority of users tweet rarely, the shorter the time periods over which behaviour is being assessed, the greater the number of values in our matrices that will be 0's. Establishing meaningful correlations when the overwhelming majority of values are 0 is near impossible.

The situation accordingly improves as we lengthen the time intervals over which behaviour is being assessed, and the matrices become less sparse. The distribution of correlation coefficients is healthier for data covering three-day intervals, although over 95% of the correlation coefficients are still insignificant. Covering week-long intervals sees another improvement with at least 10% of the 150 million correlation scores significant, suggesting that this will be the best interval to work with.

The second thing of note is the large proportion of correlation coefficients that are negative. This was not expected - although here, again, we must bear in mind that most of the values are so low as to be insignificant. Further discussion of the negative correlations appears in Section 6.3, following an assessment of community contributions to the population behaviour.

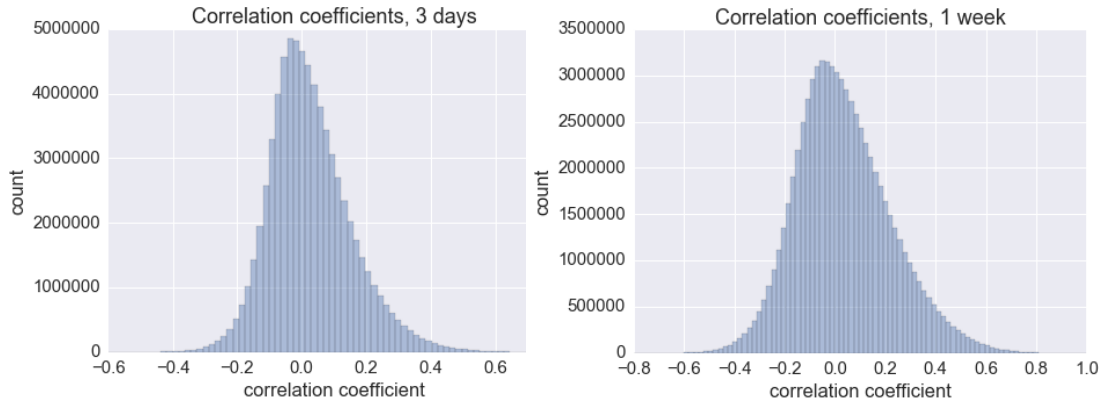


Figure 4: correlation coefficients for patterns of three-day (left) and weekly (right) behaviour.

4.1.b(ii) Data with categorical values

The same approach was taken for calculating correlation coefficients on the categorical data – where values indicated simply whether or not users had tweeted during an interval of time. While using categorical values would not replace 0's in a behaviour time series with 1's, it was expected that it would make finding correlations easier, with a wider range of correlation coefficients being returned as a result.

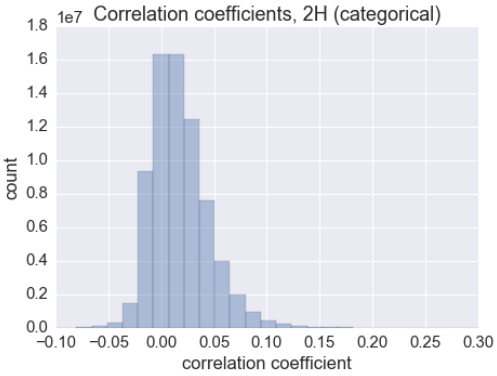


Figure 5: correlation coefficients for categorical, two-hourly behavioural data.

This did not happen, however. For two-hourly activity, a very similar distribution of correlation scores is returned (see Figure 5, right), albeit with slightly more positive correlations than before. For all the other interval lengths, though, the process of calculating correlation coefficients did not work. In each of the daily, 3-day and weekly cases, there are multiple users with 1's for every interval. Calculating correlation coefficients for these users with *pandas* returns NaN values. Although a workaround could have been found (see Section 6.3), the result for two hours was sufficiently unencouraging that we opted to press ahead using continuous-valued data alone.

4.2 Finding communities

We adopted a two-pronged approach to finding sub-communities within the population. In the first instance, we attempted a hierarchical clustering of the data. In the second, we applied a community detection algorithm to the networks resulting from thresholded correlation matrices.

4.2.a Hierarchical clustering

4.2.a(i) Explanation

Hierarchical clustering is an approach to grouping similar objects that works either from the bottom up, or the top down. The "bottom up", or "agglomerative", method is the more common of the two. It starts by placing each object – user, in our case – into its own cluster. It then uses a distance metric and a "linkage" method (see the end of 4.2.a(ii)) to evaluate how similar the clusters are to each other. The most similar two clusters at each iteration are combined, with this merging represented as one level in a hierarchy of cluster unifications. Subsequent instances of cluster combination constitute progressively higher levels, until all the objects are merged into one vast agglomeration at the top of the hierarchy. The "top down" approach works the other way, starting with one mega-cluster and progressively dividing it – it is typically referred to as "divisive" clustering – until each object occupies its own cluster.

We opted to use hierarchical clustering here primarily because it does not require the specification of a desired number of clusters before starting the process, as the commonly-used *k-means* approach does. Allowing the data to speak for itself, with any identifiable clusters returned by the algorithm, seemed preferable to projecting our preconceived expectations onto it. Hierarchical clustering is also good to work with because the hierarchy of clusters it generates can be visualised clearly as a dendrogram. Prominent clusters can be identified from these dendrograms, with sensible levels at which to "cut" the dendrogram – returning the clusters identified up to this point – easily deduced (see Section 5.1).

4.2.a(ii) Implementation

Scipy's (Jones et al., 2001) "clustering.hierarchy" suite was used for the hierarchical clustering. Although divisive clustering would also have worked, the default agglomerative approach was adopted. It was noted above that hierarchical clustering uses a distance metric to establish the pairwise similarity between points – and, subsequently, between clusters. However, we had chosen to represent the relationships between users as correlations, rather than distances (see Section 4.1.b). As such, a conversion of our correlation matrix to a distance matrix that maintained the associations between users enshrined in the original was required.

One solution was to use a measure of squared Euclidean distance. Provided the data for each user has been scaled to have 0 mean and unit standard deviation, both squared Euclidean distance and correlation reduce to the cosine rule (Cross Validated, 2015). The result is that one can be substituted for the other. We accordingly scaled the data with scikit-learn's StandardScaler and used *scipy*'s "pdist" function to find the squared Euclidean distances between users. The resulting distance matrix was passed to *scipy*'s "linkages" function, which carried out the clustering.

Hierarchical clustering was attempted for each of the four time intervals being considered: two hours, one day, three days, and one week. On the basis of the correlation matrices discussed in Section 4.1.b(i), it was expected that the data on weekly behaviour would yield the best results. Four linkage methods were also experimented with. "Single" is the default method used by *scipy*'s "linkage" function, and calculates the distance between two clusters according to how far apart their nearest points are. For the "complete" approach, the points in the clusters that are furthest apart are used instead. Finally, adopting the "average" method results in a calculation based on the average position of each cluster. The results are presented in Section 5.1.

4.2.b Community detection

An alternative way of finding clusters of similar users was to create networks from our correlation matrices. Running a community detection algorithm on these networks would identify clusters of similar users on the basis of their higher interconnectedness when compared to the graph as a whole.

4.2.b(i) Explanation

As with more typical clustering algorithms, a range of different possible approaches to community detection exist. We opted to use the 'Louvain Method', primarily because it has been proven to work quickly on large graphs. While ours would consist of only 11982 nodes, we hoped to run the community detection process many times over – and often on networks where the degree of the nodes was likely to be very high (see 4.2.b(i)). The Louvain Method has also been used on similar problems with success (e.g. Bagheri et al., 2015), and was itself originally employed to assess the integration of a social network (Blondel et al., 2008).

The Louvain Method is similar to agglomerative hierarchical clustering in that it begins by assigning each node to its own community. Nodes are moved to new communities to maximise the modularity of the network. Then a new network is constructed, with the clusters from before used as nodes. The process repeats, and this happens until there are no further changes in community membership.

4.2.b(ii) Implementation

For community detection to work, a network is required. Our approach was to threshold one of our correlation matrices, producing an adjacency matrix which could then be used to generate a graph. The matrix of correlations based on weekly time periods was used, since it contained by far the largest proportion of meaningful correlation scores out of those considered (see 4.1.b(i)). We used 50 different threshold values spread evenly between 0 and 0.8, employing scikit-learn's Binarizer to convert the correlation coefficients to 1's or 0's, and passed the resulting adjacency matrices to *NetworkX*'s (Hagberg et al., 2008) "Graph" function. These graphs were then passed to Thomas Aynaud's (2009) implementation of the Louvain algorithm for the community detection process.

Since most of the correlation coefficients were clustered around 0 even for the weekly data, the lower the threshold, the more 1's there were in the adjacency matrix. The nodes (users, for our purposes) of the resulting graphs were accordingly of higher degree, as can be seen in the distributions in Figure 7 (page 16). Applying a threshold of 0 to the correlation matrix results in a large number of nodes being connected to *most* of the other nodes (Figure 7, A). By contrast, higher threshold values cut the number of edges between nodes dramatically. The distribution even for a threshold of 0.5 (Figure 7, F) looks to follow a power law, with most nodes of low degree and a long tail of progressively fewer nodes of high degree. This power law distribution becomes even more pronounced as thresholds are increased yet further. Since power-law distributions are often plotted on doubly logarithmic axes, the results for thresholds of 0.6 and 0.7 are presented thus (Figure 7, G and H).

Percolation theory (Grimmett, 1999) would suggest that there is a threshold beyond which one giant connected component exists equal to the size of the system – guaranteeing, in other words, that from any node in the network, there

exists a path to any other node. In our case, this threshold should correspond with a threshold applied to our weekly correlation matrix, below which connectivity between nodes is so great that one giant component exists, and above which lower levels of connectivity result in a network consisting of multiple connected components.

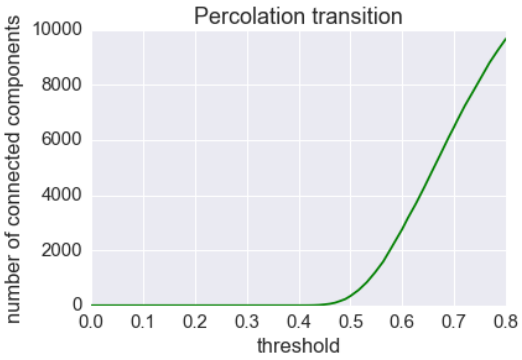


Figure 6: the impact of changes to threshold on the number of connectied components in the resulting graphs.

This expected pattern of percolation is visible in our data (see Figure 6, right). A giant connected component is found for threshold values below a point around 0.42, with the number of connected components rising rapidly after that. Ultimately, each node constitutes its own connected component.

Intuitively, there should exist a correlation between this and the number of communities found by the Louvain method. For graphs with one giant connected component, the number of communities detected will approach 1 – although more than one community may still be found within the giant connected component, particularly at higher threshold levels where the connections between nodes remain more sparse. A network with multiple connected components, meanwhile, should necessarily have at least as many distinct communities as it has connected components. Once we have surpassed the percolation threshold, then, the number of communities should also rise rapidly with the threshold level.

Percolation theory has been extended to cover the formation of cliques in random networks, with one giant clique the inevitable result once a certain probability of two nodes being connected is applied (Palla et al., 2006). This could be seen to apply in our case again, with a lowering of the threshold at which the correlation matrix has been binarised increasing the probability of nodes being linked – albeit not randomly. Taking the above observation about the likely relationship between the number of connected components in the network and the number of communities detected, we anticipated a clique percolation threshold to exist somewhere close to and probably just below the standard percolation threshold.

On this basis, the best value to use for thresholding the correlation matrix ought to be something close to the percolation threshold: significantly below, and the population as a whole would likely be returned as a single community; significantly above, and too many communities would be returned, fast approximating to the behaviour of individuals and increasingly meaningless as a result. The relationship between the number of connected components, the number of detected communities and their sizes, and the modularity of the community partitions is discussed in Section 5.2.

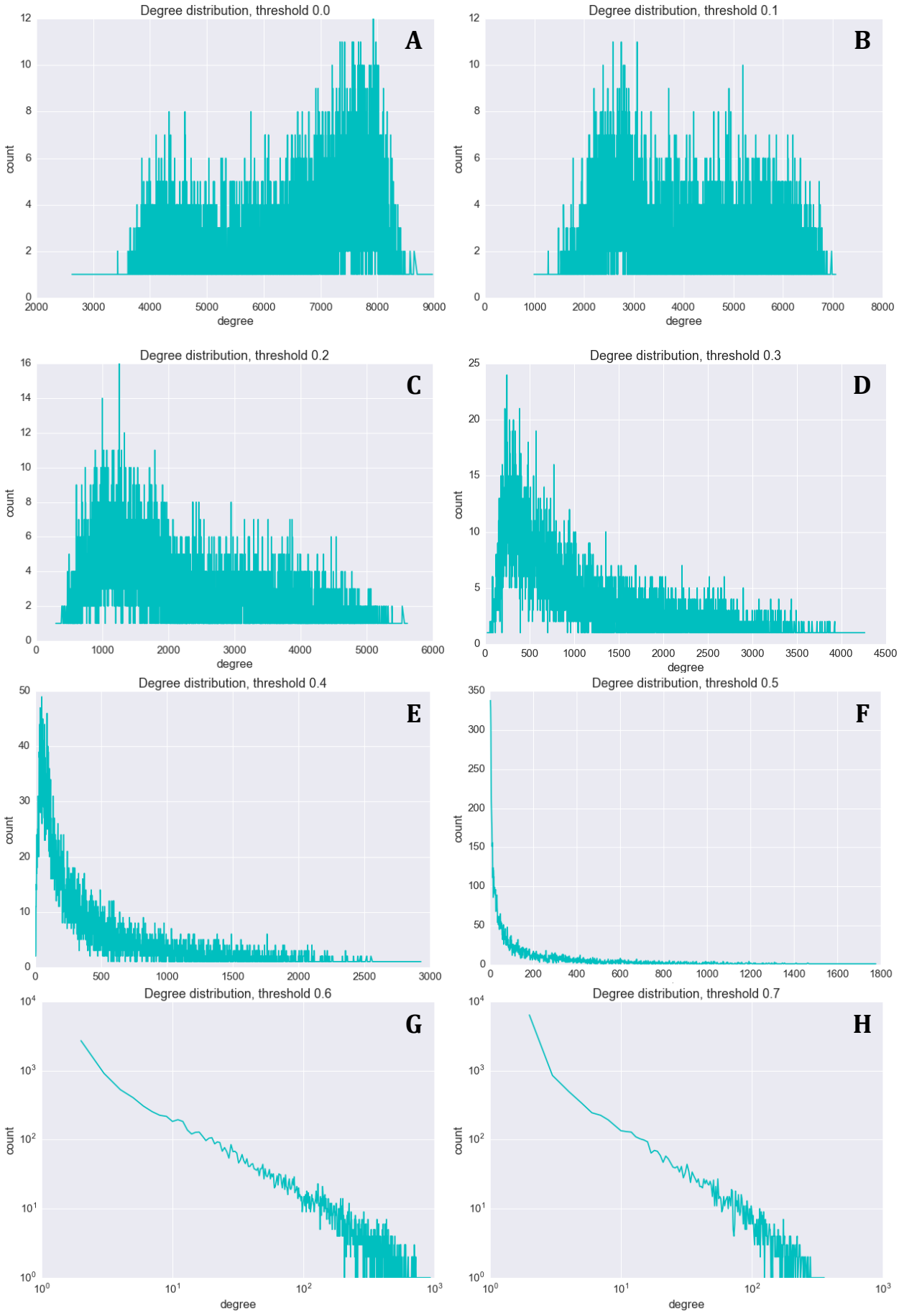


Figure 7: degree distributions for graphs created by thresholding the weekly correlation matrix at different levels.

5. Results

5.1 Hierarchical clustering

The results of hierarchical clustering were assessed on two primary bases. In the first instance, the cophenetic correlation between the linkages found for the data and the original distance matrices was calculated in each case. This was used as a measure of the fidelity of the clusters to the original correlations between users.

However, given that the majority of the original correlation values were deemed insignificant and likely to lead to poor clustering results (see Section 4.1.b), a more detailed visual assessment of the clustering behaviour was also necessary. Dendrograms were plotted for this. It was possible that the cophenetic correlations could be high without the clusters being at all distinct or informative – particularly for data covering shorter time intervals. As such, strong cophenetic correlations should be seen as further validating rather than defining apparently meaningful results.

5.1.a Cophenetic correlations

Interval length	Linkage metric	Cophenetic correlation
1 week	Single	0.162666
1 week	Complete	0.421130
1 week	Average	**0.514487**
3 days	Single	0.147632
3 days	Complete	0.304161
3 days	Average	**0.469375**
1 day	Single	0.144060
1 day	Complete	0.277940
1 day	Average	**0.421593**
2 hours	Single	0.231390
2 hours	Complete	0.123423
2 hours	Average	**0.496110**

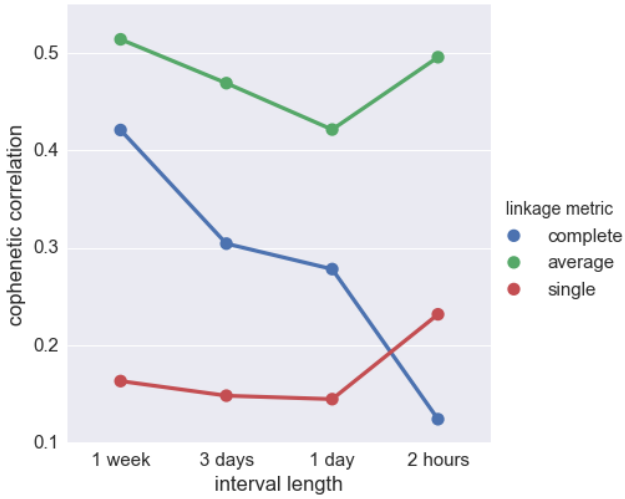


Table 2, and **Figure 8**: comparing the impact of interval length and clustering method on cophenetic correlation scores.

The table and graph above (Table 2, Figure 8) demonstrate that average clustering is the most faithful to the original correlations found between users in every case, and that those relationships are generally best maintained for the correlation matrix in which they were strongest – that representing weekly behaviour. It is interesting that the relationships in the correlation matrix for two-hourly behaviour, which almost exclusively contained values below a meaningful threshold, are better maintained by single and average clustering than is the case for the other time intervals (with the exception of the weekly, average clustering case). The trend evident for complete clustering is much closer to what was expected, and looks as though it might approximate to a straight line if the gaps between time intervals were more consistent.

5.1.b Visualising clustering behaviour

Further to the low cophenetic correlations resulting from the 'single' clustering method, at every level of temporal granularity no distinct clusters of users within the population were found. This is demonstrated in Figure 9 – a visualisation of the last 30 clusters created for the weekly data – which shows no discernible structure, with users grouped together in one ever-larger, seemingly homogenous mass. The same pattern of activity is seen in dendrograms for the other interval lengths.

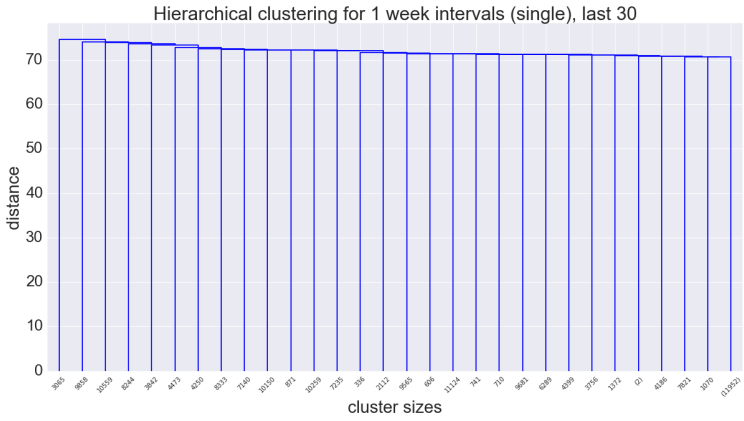


Figure 9: dendrogram from single hierarchical clustering on weekly data.

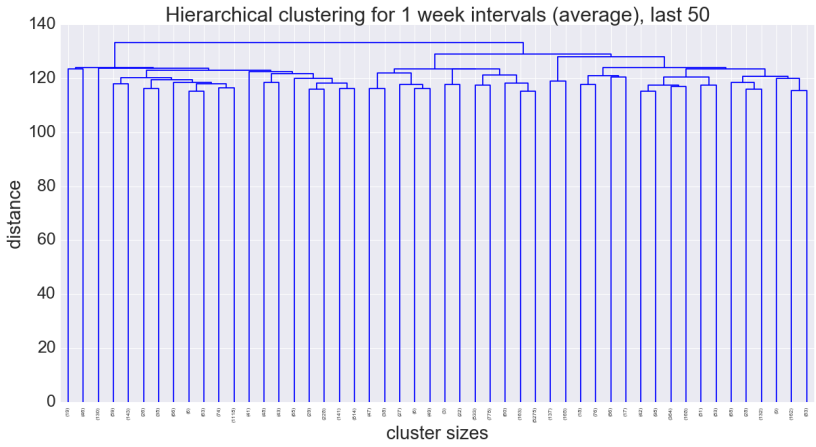


Figure 10: dendrogram from average hierarchical clustering on weekly data.

Adopting the average clustering method sees the situation improve. Some structure begins to appear for the one-week (see Figure 10) and three day intervals, although the clusters at higher levels of temporal granularity remain poorly defined.

Complete clustering reveals a much more distinct structure for every time interval – including (albeit less successfully) two hours. Focusing again on the weekly data, since the correlations here were the most significant, three large, distinct clusters that would be returned by cutting the dendrogram at a distance of about 225 are visible: one occupying each of the left, right, and centre of the dendrogram. Moving further down the hierarchy, some 4-6 slightly smaller but still distinct clusters would be returned by cutting at a distance a shade above 200, below which the population begins to splinter more rapidly. Considering the behaviour of users in some of these larger clusters would seem a sensible starting point for assessing how effectively the clustering process has differentiated between groups with distinct patterns of activity. See Figure 11 (page 19) for dendrograms covering behaviour on data from 1 week, 3 day and 1 day intervals, and Figure 12 (page 20) for two hour data.

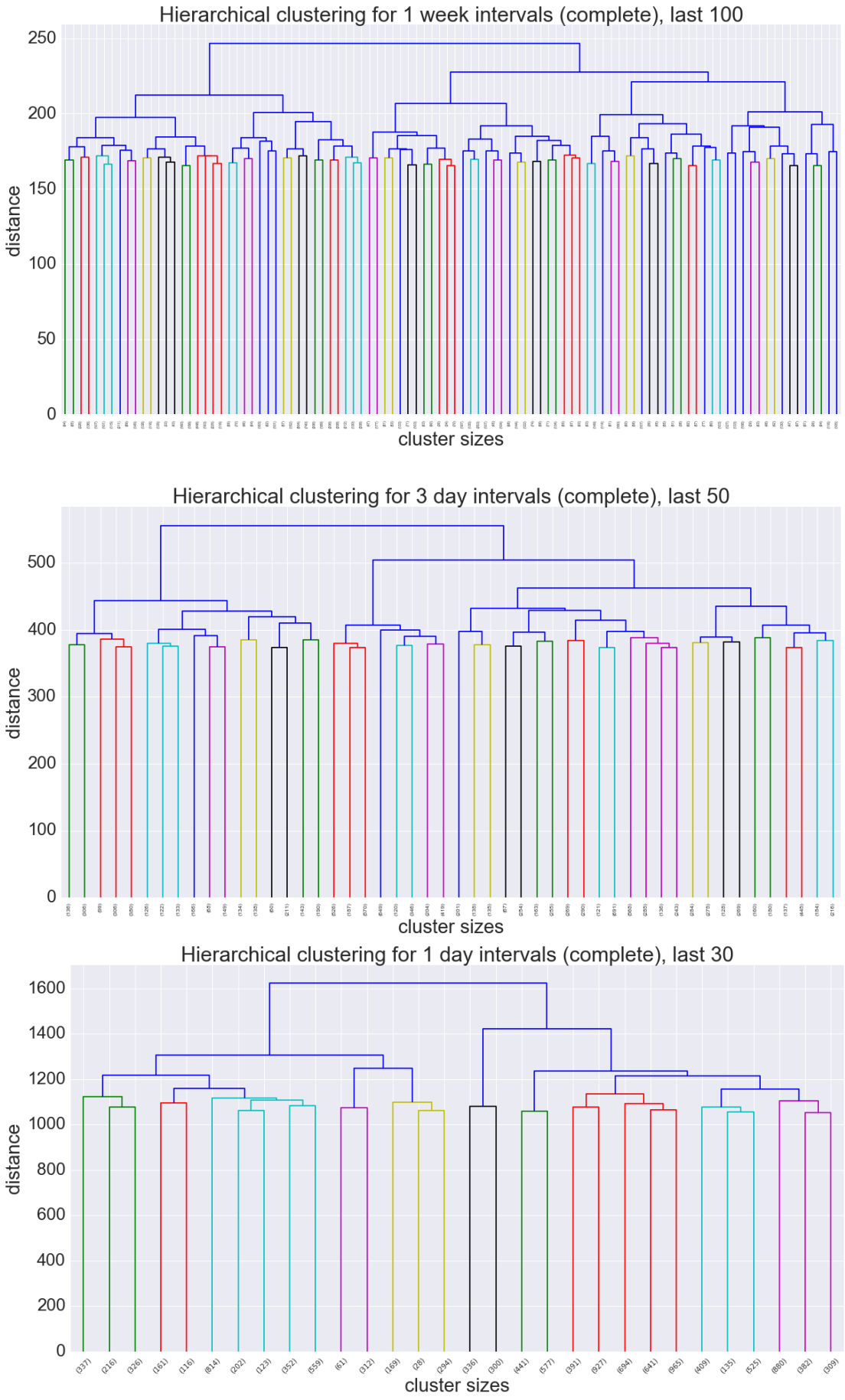


Figure 11: dendrograms from complete hierarchical clustering on data for 1 week (top), 3 day (middle) and 1 day (bottom) intervals.

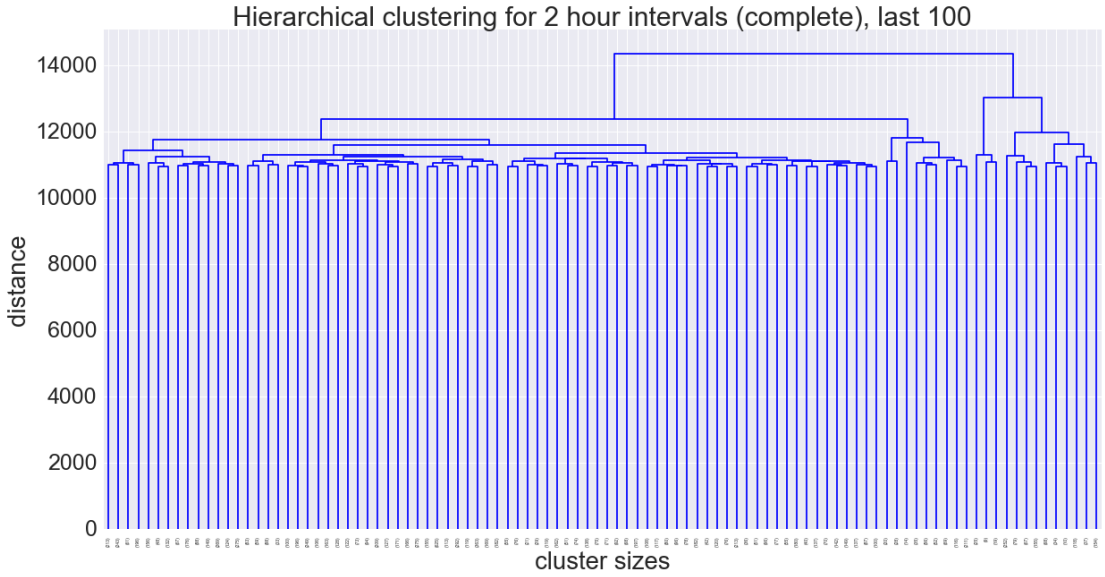


Figure 12: dendrogram from complete hierarchical clustering on data for 2 hour intervals.

5.2 Community detection

Community detection results are typically assessed by consideration of how far each partition of the graph has maximised modularity. Our results showed a steady increase in modularity score with the threshold used to create the corresponding adjacency matrix, until a kick up towards the end as the number of communities returned approximated to the number of nodes in the network. While the score climbed to 0.5, this was for partitions where almost every node was its own community, which is not hugely helpful. We consequently focused on how distinct the behavioural patters of communities were (see 5.3) rather than modularity for assessing the performance of community detection here.

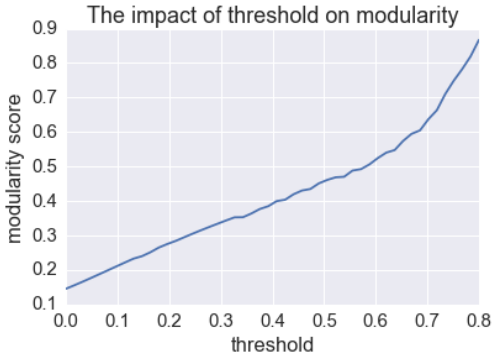


Figure 13: visualising the increase in network modularity with threshold values.

Unexpectedly, our results do not show a clear clique percolation threshold. No fewer than 3 communities are detected all the way down to thresholds of 0.0. It was expected that the number of communities detected would drop more quickly to 1, given that the standard percolation threshold – below which the network consists of one giant connected component – is around 0.42. This is likely a reflection of the number of negative correlation values in the original matrix: reducing the threshold yet further, to negative values, would apparently be required to identify one giant community. For the purposes of trying to identify meaningful communities, though, that there remain multiple distinct

communities even once the standard percolation threshold has been reached can be seen as positive.

Our intuitions (see Section 4.2.b) about the relationship between the percolation threshold of our graphs and the communities detected within them by the Louvain method were largely validated though, with the number of communities detected closely following the number of connected components. Prior to hitting the percolation threshold, the number of communities was consistently low and seemingly approaching 1 for lower thresholds – the clique percolation threshold we were expecting to reach. Above the standard percolation threshold, the number of communities increases rapidly, ultimately approximating to the number of nodes in the network.

This splintering of communities from the percolation threshold onwards manifests itself primarily as single nodes breaking off to form their own, isolated communities. Between thresholds of 0.42 and 0.46, the number of communities that consist of a single node jumps from none to over half of the total, and the proportion continues to rise from there, surpassing 90% by thresholds of 0.57. The complementary observation to make is that 4-6 communities – the same number of communities as identified just before the percolation threshold – retain membership numbers in the thousands at thresholds as high as 0.64. The consistent detection of roughly the same number of large communities as the network around them splinters suggests that these communities should be both stable and meaningful.

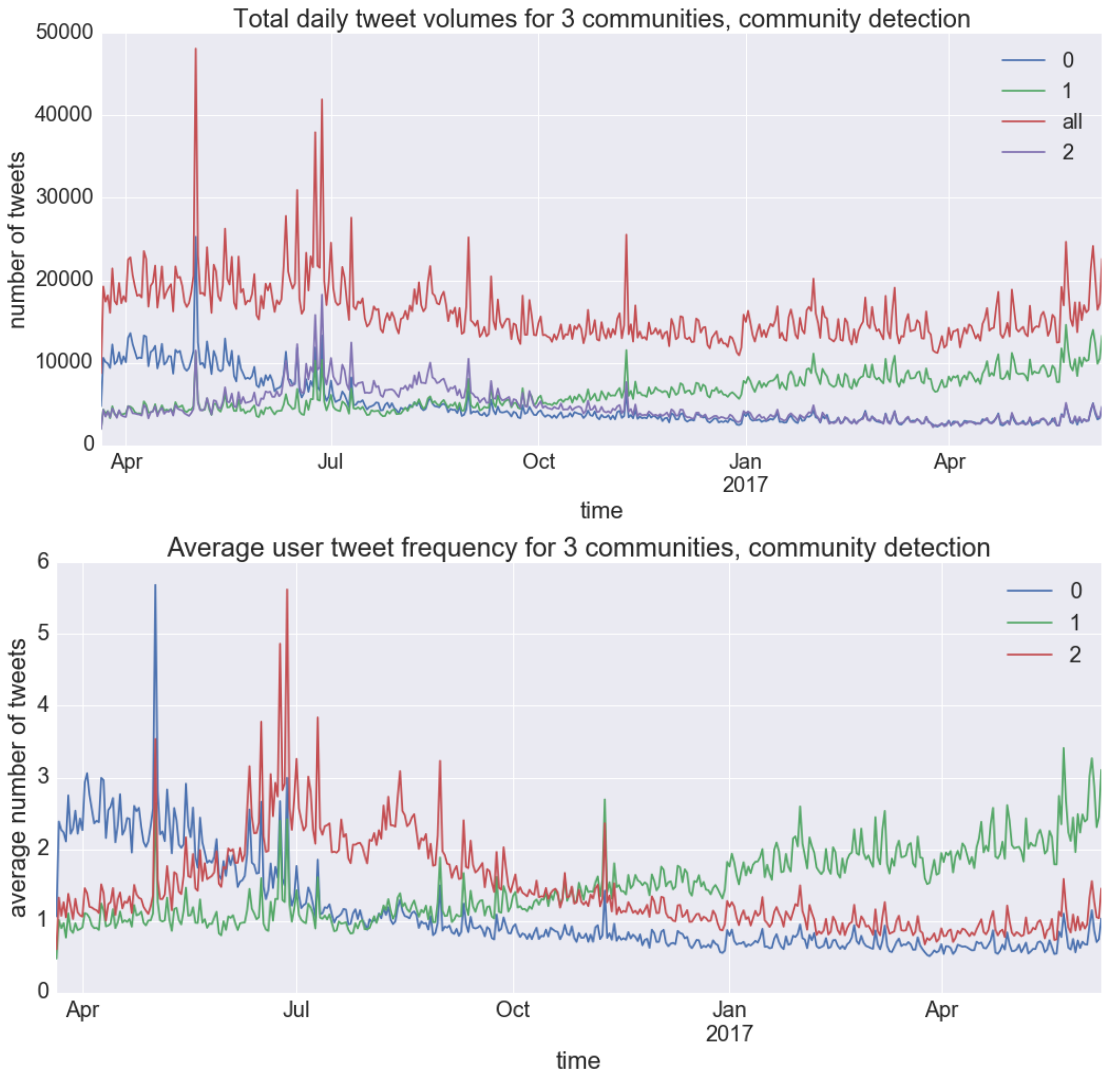
5.3 Visualising community contributions

The results of the hierarchical clustering and community detection processes were used to plot the behavioural profiles of identified sub-communities against each other and against that of the population as a whole. In the first instance, total daily tweet volumes over time were plotted for each community and the population. Secondly, average daily tweet counts for members of the respective communities were plotted. The plots of total volumes are intended primarily to compare the contributions of each community to the overall population's tweet behaviour. The plots of average tweet counts are intended to enable a more detailed comparison of the communities against each other – some communities are too small for patterns in their activity to be easily identified from the volume plots – and population behaviour is accordingly not included.

We focused on assessing the activity of between 3 and 6 communities, based both on our observations further to the results of community detection (see the previous section, 5.2), and on the fact that cluster numbers in this range also seemed easily identifiable in our hierarchical clustering results. Despite the communities being identified based on weekly behaviour, we found plots over daily time intervals to be more interesting and informative, and these are used accordingly. The communities are ordered according to size, with community 0 corresponding to the largest in the given example, then community 1, and so on. The plots appear over the course of the following pages, with accompanying interpretations.

5.3.a Three communities

5.3.a(i) Community detection



Community	Size
0	4447
1	4286
2	3249

Figure 14: plotting the behaviour of the 3 communities identified by community detection at a threshold of 0.08. (Top) volume of daily tweets from community and population; (bottom) average number of tweets per person in communities; (left) community sizes.

Even where the original correlation matrix was thresholded at a very low level, community detection on the resulting graphs resulted in the identification of at least 3 distinct communities. Figure 14 illustrates the behaviour of the three communities retrieved at a threshold of 0.08.

There are a number of things to note here. The first is the striking divergence in the patterns of activity for community 0 and community 1. Community 0, despite constituting little more than one third of the population, appears to account for over half of the total population activity at the start of the period in question. However, community 0 becomes rapidly less active through May, June and July of 2016. That decline steadies, but continues throughout the period covered, with

community 0 ultimately contributing only around one quarter of the total population's tweets by June 2017.

The behaviour exhibited by community 1 is astonishingly similar to that of community 0, but in reverse. In April 2016, members of community 1 account for no more than one quarter of the total population's tweet volume; 14 months later, the same community appears to account for nearly three quarters of the total volume. This mirroring of the two communities might reflect the preponderance of negative correlation values in the original correlation matrices.

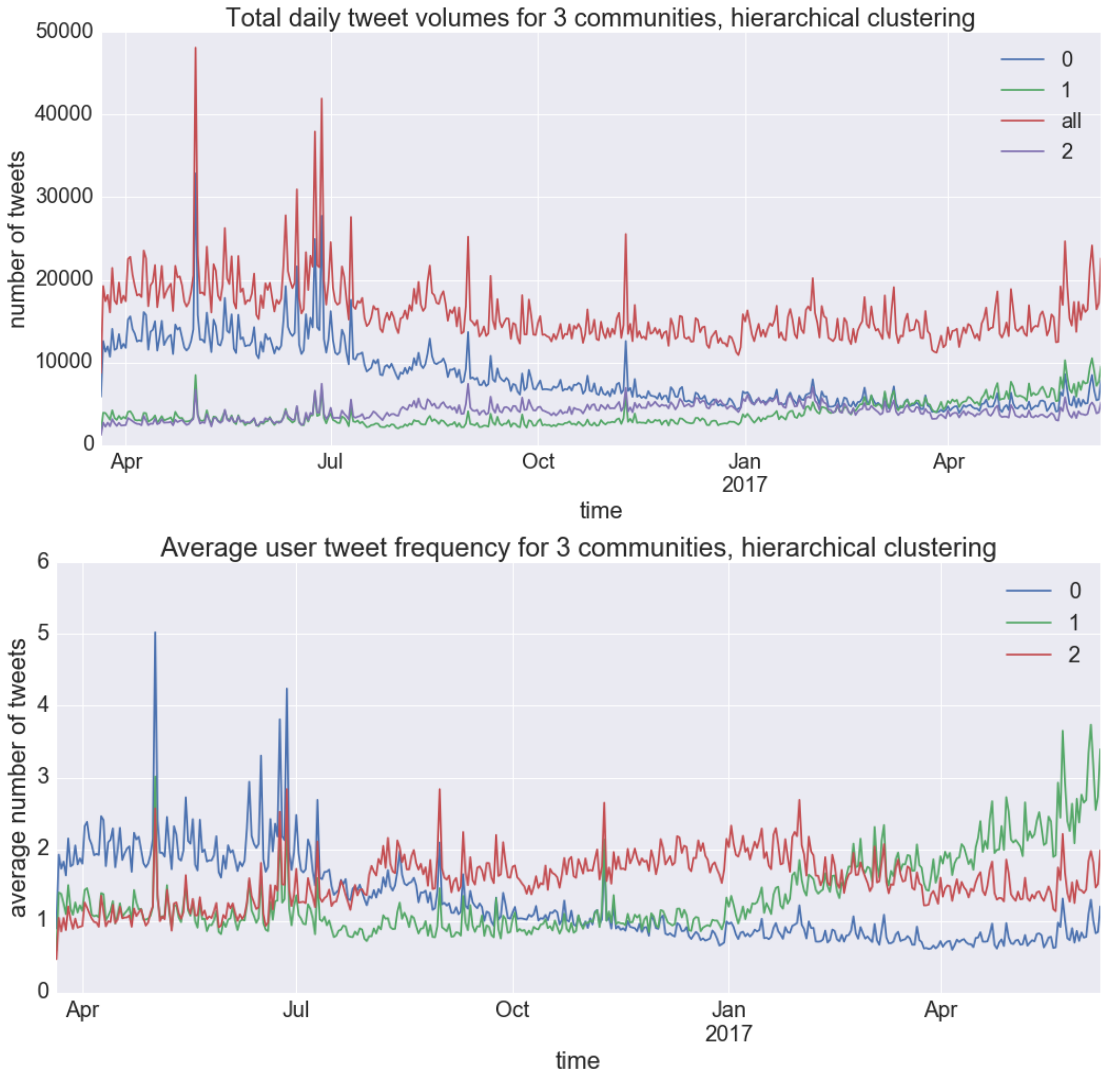
The activity of community 2 also has a distinct peak, although this occurs when the behavioural patterns of communities 0 and 1 cross over rather than at one end of the time period. The 'off-peak' behaviour of this community is also very similar to that of communities 0 and 1, with smaller, daily peaks matching very closely with those of the other communities (see the bottom left and right of the top plot). Members appear to tweet slightly more on average during these down periods than those of the other communities, and the proportion of population activity accounted for by community 2 during these times is consequently (given this population is smaller) very similar to that of communities 0 and 1 during their respective lulls, at around one quarter.

5.3.a(ii) Hierarchical clustering

There is a clear correspondence between the behavioural profiles of the 3 communities identified by community detection and those returned by hierarchical clustering, as can be seen by comparing Figures 14 (page 22) and 15 (page 24). Again, the activity of one peaks at the beginning before declining, another peaks at the end after a slow start, and the third peaks somewhere in the middle where the activity of the two others has concurrently dropped off.

However, there are also clear differences. Community 0 is much larger than either community 1 or 2 here; community detection returned communities of much more consistent sizes. The activity patterns visualised below suggest that while users placed in community 0 by the community detection approach have been allocated similarly by hierarchical clustering, most of those previously in community 2 have also been placed community 0. The otherwise depleted numbers in community 2 are bolstered by the addition of a number of users assigned to community 1 by community detection. As a result, the peak in community 0's activity covers a longer period and has a slower drop-off, claiming most of what was community 2's peak in the community detection example. While the activity of community 1 still reaches its zenith towards the end of the 14-month period, the build-up to this point is less protracted, happening over the course of 6 months rather than 10. Community 2's peak, meanwhile, is less pronounced, and takes place further into the period – towards the end of 2016, rather than during the summer.

While these results suggest that distinct communities can be identified, they remain too coarse to realistically infer what those communities might represent.



Community	Size
0	6547
1	2815
2	2620

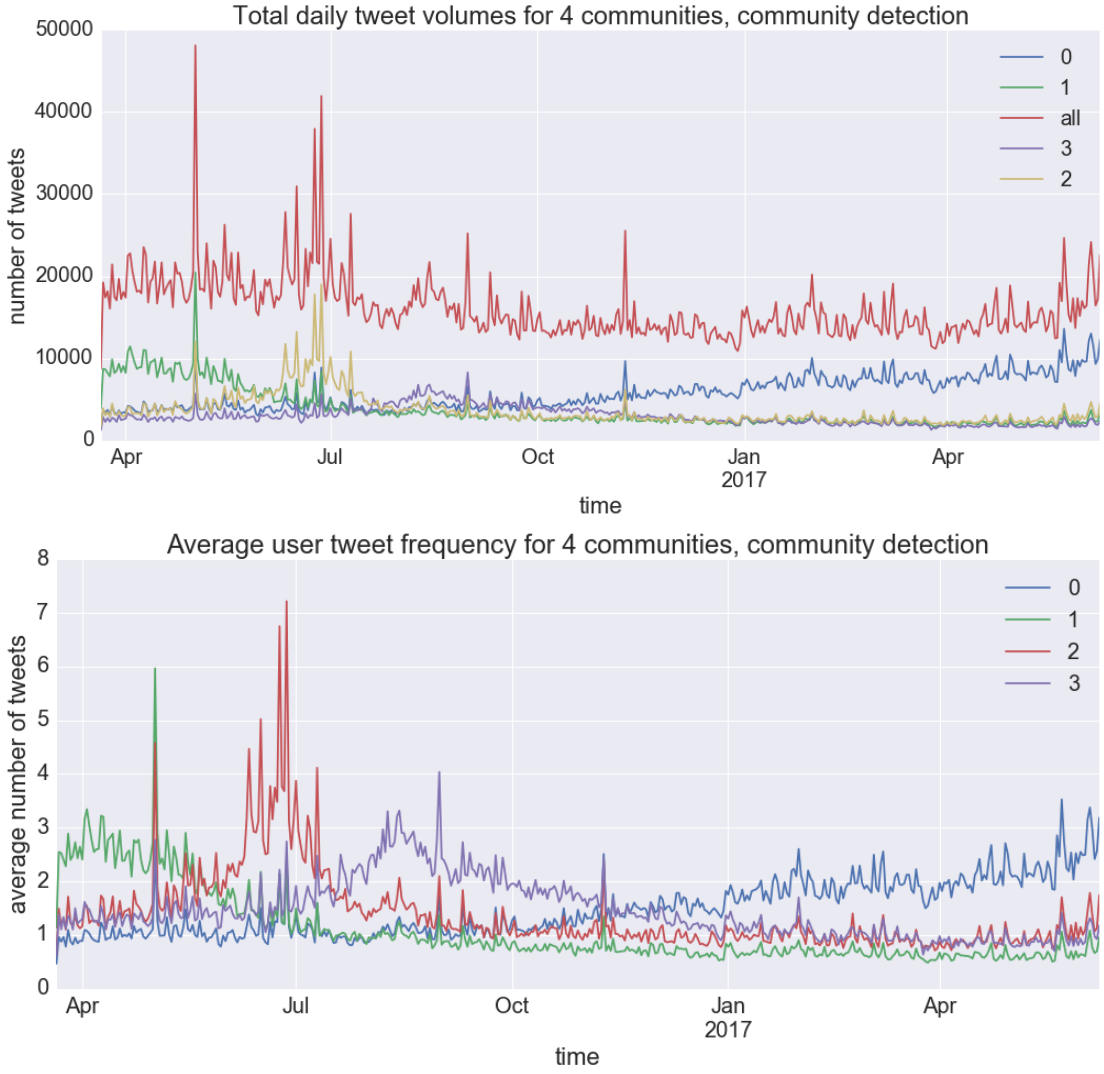
Figure 15: plotting the behaviour of the top 3 clusters retrieved by complete hierarchical clustering. (Top) volume of daily tweets from community and population; (bottom) average number of tweets per person in communities; (left) community sizes.

5.3.b Four communities

5.3.b(i) Community detection

Four communities were typically returned by community detection for thresholds between 0.15 and 0.35. The behavioural patterns evident for the three communities returned at lower thresholds are generally maintained, as can be seen in the example for a threshold of 0.28 below (Figure 16). As before, two communities' activity patterns peak at either end of the time period. What was previously the peak for the third community is still apparent, but split between two this time, with the month-long peak of one covering most of June and the beginning of July 2016, while that of the other roughly spans August and September.

At threshold levels corresponding to the transition between 3 and 4 communities, or 4 and 5, the sizes of the communities become more inconsistent. But for the example below, placed between these two points, community sizes remain fairly even. It is worth noting, though, that the largest community, community 0, is no longer the one that peaks at the start, but the one that peaks at the end of the period. Both have decreased in size, contributing members to the new community.



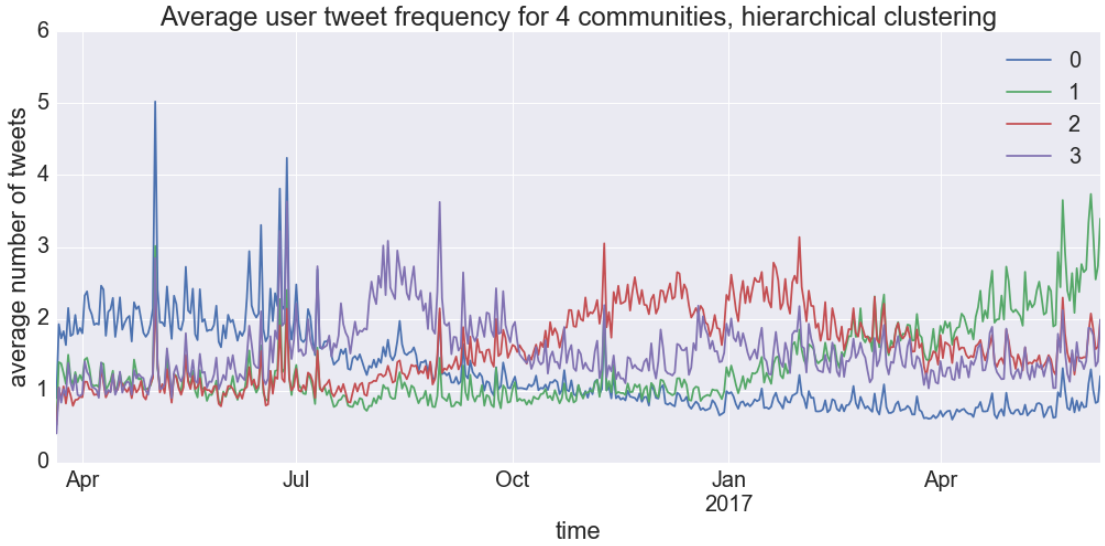
Community	Size
0	3862
1	3428
2	2634
3	2058

Figure 16: plotting the behaviour of the 4 communities identified by community detection at a threshold of 0.28. (Top) volume of daily tweets from community and population; (bottom) average number of tweets per person in communities; (left) community sizes.

5.3.b(ii) Hierarchical clustering

Retrieving four communities from the hierarchical clustering process returns the same result as for three, except for the fact that one of the communities is split into two, as would be expected. The profiles for the largest two communities remain the same in this case, while the smallest community roughly halves in

size and shares the peak in the middle of the period with the new addition. As with the community detection case, a distinct community-specific peak now occurs across the months of August and September. However, the peak for the fourth community takes place over the end of 2016 and beginning of 2017, rather than over the months of June and July 2016. As before, then, all but one of the communities returned by community detection and hierarchical clustering exhibit similar behavioural profiles.



Community	Size
0	6547
1	2815
2	1396
3	1224

Figure 17: plotting the behaviour of top 4 clusters retrieved by complete hierarchical clustering. (Above) average number of tweets per person in communities; (left) community sizes.

5.3.c Five and six communities

Communities become more distinct once five or six are being identified. Although similar patterns persist for both community detection and hierarchical clustering, community-specific peaks that were previously prolonged or not especially pronounced begin to be split between multiple communities.

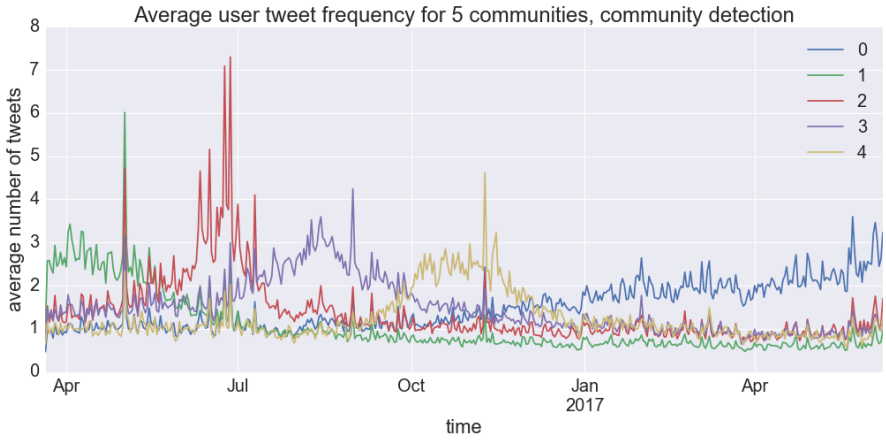
5.3.c(i) Community detection

Comparing Figure 16 with Figure 18 demonstrates this well. Where before community 3 was seen to have a peak covering August and September 2016 that dragged on, albeit less prominently, into October and November, once an additional community is added this becomes better defined. Community 3's activity still peaks around August, but the uptick in activity from October towards the end of the year is now attributed to community 4.

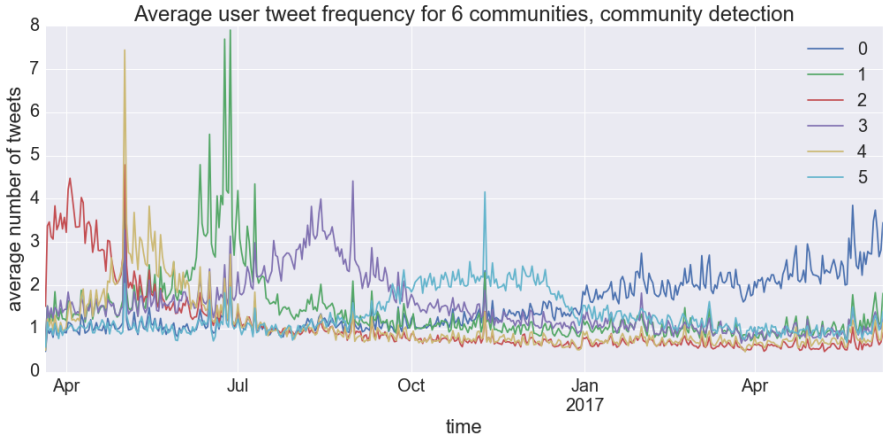
A similar split is evident moving from a consideration of five communities to six. Here, the community that was particularly active at the start of the period is divided into two. There remains an initial peak over the end of March and start of

April 2016, but this drops off rapidly, with a second peak attributed to a new community covering May.

Note that the 6 communities retrieved here were actually from a threshold level that returned a total of 29. It is only at these higher levels that six communities of a consistent size exist – when only six are returned, one community only has three members. When communities are that small, it is both difficult to see their contributions to the overall volume of the population's tweets, and hard to assess how meaningful peaks in their average daily tweet count are, since the average may simply be driven up by one suddenly active user.



Comm.	Size
0	3734
1	3152
2	2516
3	2061
4	519



Comm.	Size
0	3239
1	2139
2	2007
3	1823
4	1597
5	1049

Figure 18: average daily tweet frequency for five and six communities retrieved by community detection.

5.3.c(ii) Hierarchical clustering

In the case of hierarchical clustering, the peak in activity over June and July is attributed to one community once 5 communities are considered, with the largest community finally splitting. The activity for a 6th cluster, once added, is less clearly defined, but appears to have a peak around February and March of 2017 – a period during which community detection sees one group as having a near-monopoly. This suggests that combining the two approaches might be helpful for emphasising different aspects of the community structure in the population.

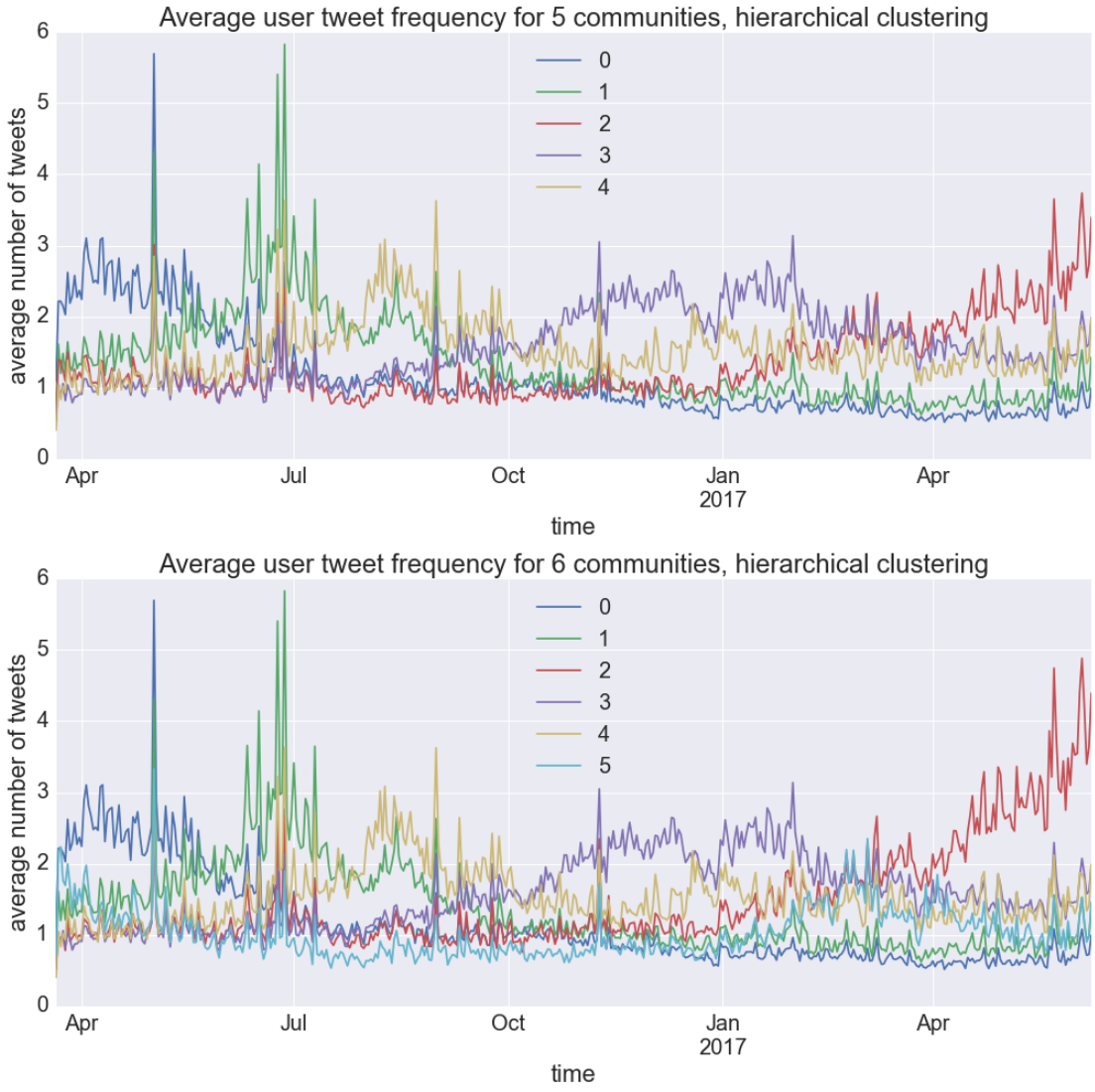


Figure 19: average daily tweet frequency for five and six communities retrieved by hierarchical clustering.

5.3d Interpreting community contributions

At this point, it seems reasonable to attempt to identify what the detected communities represent. We focus on the case of community detection that returned five communities, since the boundaries between community activity peaks here seem particularly distinct. The average per-user daily behaviour of these communities is plotted above (Figure 18, top), with the total daily tweet volumes for each plotted against those of the population as a whole on page 32 (Figure 20). We will work through the figures from left to right.

5.3.d(i) Leicester fans

The first visible community (community 1, green), distinctive for its surge in activity at the start of the period and subsequent decline, has been evident throughout the process across both the community detection and hierarchical clustering approaches. We interpret this community as consisting of fans of Leicester City football club.

This might sound strange, given that the entire population is made up of followers of the Leicester City's official Twitter feed. However, analysis of match day tweet activity indicates that most followers of Leicester City are interested observers – often fans of other Premier League football clubs, and users interested in football or sport more generally – rather than fans of the club itself. Even many followers that *are* fans might be considered observers: interested in following the club's fortunes, without necessarily feeling sufficiently strongly to tweet about them.

We infer that this community constitutes Leicester fans, rather than mere observers, for multiple reasons. Firstly, there is a pronounced peak in user activity for this community on the 2nd May 2016, when Leicester City won the Premier League. Leicester City winning the League was widely considered to be one of the most astonishing outcomes in football, and perhaps even sporting, history. As a result, it would not be unreasonable to expect every active follower of the club to tweet about it – fan or otherwise – and a similar peak is accordingly seen for all of the communities in the population. However, the community in question stands out for the high levels of activity it exhibits in the weeks approaching that peak.

Where for many of the other communities the peak appears almost out of the blue, for this community is anticipated by considerably higher than average activity, and particularly on days when Leicester won matches (3rd and 10th April) to move progressively closer to clinching the title. It is interesting that retrieving six communities from the community detection process splits the community to show evidence of a core group of fans who demonstrate this anticipatory behaviour, and another group (community 4 (yellow) in Figure 18, bottom) that contributes little to activity before the event, but is buoyant in its aftermath.

The activity of this community (or that of both, if taking into account its subsequent division) suffers a notable decline after the end of the football season. Some peaks in activity coinciding with England matches in Euro 2016 are apparent towards the end of June, but otherwise there is little of note over that summer. Come the start of the new season in mid-August there is a brief spike in activity, with another coinciding with the transfer deadline day at the end of the summer. However, activity appears to drop steadily as Leicester perform poorly from the start of the season and into the New Year.

5.3.d(ii) (English) Football fans

Another familiar behavioural pattern, apparent throughout the community detection process and for when five or six communities are returned by hierarchical clustering, is seen for community 2 (red). Here, a surge in activity is seen over June and the beginning of July 2016, with dramatic peaks in that activity coinciding with matches played by England in the Euro 2016. While a peak is also apparent coinciding with Leicester winning the League, this community does not exhibit the high degree of anticipatory behaviour noted for the proposed community of Leicester fans. Nonetheless, sub-peaks in its activity

towards the end of the 2015-16 season and moving into the 2016-17 season do correspond with days and weekends on which Premier League matches are played and, again, there are visible spikes for the start of the new season and the closing of the transfer window. The combination of consistent responses to football matches, and particularly pronounced peaks coinciding with England matches, makes it reasonable to infer that this community is primarily made up of football fans that are English, rather than foreign, and not necessarily supporters of Leicester City.

It is worth noting that, particularly for this community, it is difficult to disentangle how far the surge in activity towards the end of June reflects responses to Euro 2016 or the build up to the EU referendum. While the beginning of this surge coincides with the start of Euro 2016, the largest peak in activity is on the day of the EU referendum result. Furthermore, the second largest peak – and third largest for the population as a whole – happens on the day of an unremarkable 0-0 draw between England and Slovakia. It is possible that there are two groups within this community: one interested primarily in football, and another in the referendum. Any apparently football-related peaks may therefore look larger than they are by virtue of being bolstered by the increasingly frenetic tweeting of the second group in the run-up to the referendum. Alternatively, it may be that there is a large crossover between England football fans and tweeters vocal about the referendum. More fine-grained analysis would be required to say for sure.

5.3.d(iii) Student sport fans

The activity of the third visible community (community 3, purple) begins to build noticeably through July and into August of 2016. This uptick appears to correspond with Wimbledon (which was won by Andy Murray) and the Rio Olympics, and initial analysis of the tweet content for the community over this period confirms that these topics do receive a lot of attention. Further football-related peaks are again apparent for the start of the 2016-17 season in mid-August, and the end of the transfer window. During the season itself, however, activity is less exaggerated.

This combination suggests two things. The first is that the members of this community appear to be sport fans more broadly, rather than fans of football alone (as might be said for the first community discussed, and the second to some extent). The second obvious inference is that this community is made up of students. While the start of the surge in activity can be seen to correspond with Wimbledon, the beginnings of it are apparent from early June – around the time that exams typically end, with school or university breaking up shortly afterwards. Similarly, the drop in activity during the football season, despite the interest shown in both its start and concurrent transfer activity, might be explained for many in the community by a return to school. Without further, detailed analysis, it would seem sensible to conclude that this community consists primarily of student sport fans.

5.3.d(iv) American "soccer" fans

The fourth visible community is by far the smallest, with members in the hundreds rather than thousands. Its pattern of activity is also especially peculiar. While an interest in football – Leicester winning the League, the start of the new season – is exhibited, and the expected spike for Brexit is still present, the main period of activity comes during a time when the behaviour of the other communities is going through a lull. This period is also very clearly defined, with a rapid increase in average per-user activity through September and into October peaking dramatically towards the start of November before quickly dropping off again.

That period of activity appears to relate primarily to the build up and result of the US Presidential Election. The result of that election on the 8th November 2016, with Donald Trump the perhaps unexpected victor, is responded to by all communities. However, much as one community exhibited activity in anticipation of Leicester winning the League, the only community that really seems to respond the events leading up to the election, with debates between the candidates running through October, is this community. This otherwise unusual emphasis on US rather than UK events suggests that this smaller community consists primarily of football (or "soccer") fans living in America.

5.3.d(v) The potentially politically-aware

The final visible community is more difficult to define. Evident, again, throughout the community detection process, it starts the period relatively inactive, but ends it contributing the majority of the overall population's tweets.

What does seem clear is that its members are primarily British. Particularly when compared to the likely Americans discussed above, they respond very strongly to the terror attacks in London and Manchester that took place in March, May and June 2017. The long build-up in activity also seems to lead to the General Election of June 2017, indicating – in combination with pronounced responses to both Brexit and the Trump election, and typically smaller peaks for sport-related events when compared to other communities – that this community might be interested primarily in political happenings.

This is the most tenuous of the interpretations, though, and the community should ideally be divided further. The politically-aware profile suggested above fits better, for instance, with community 2 retrieved from hierarchical clustering to return six communities (see the red line in Figure 19, bottom). In this result, in fact, activity over the New Year and around the time of Trump's inauguration is attributed to our proposed community of Americans, which would make more sense (see the purple line in Figure 19). This suggests that combining approaches to identifying communities could lead to more definitive results.

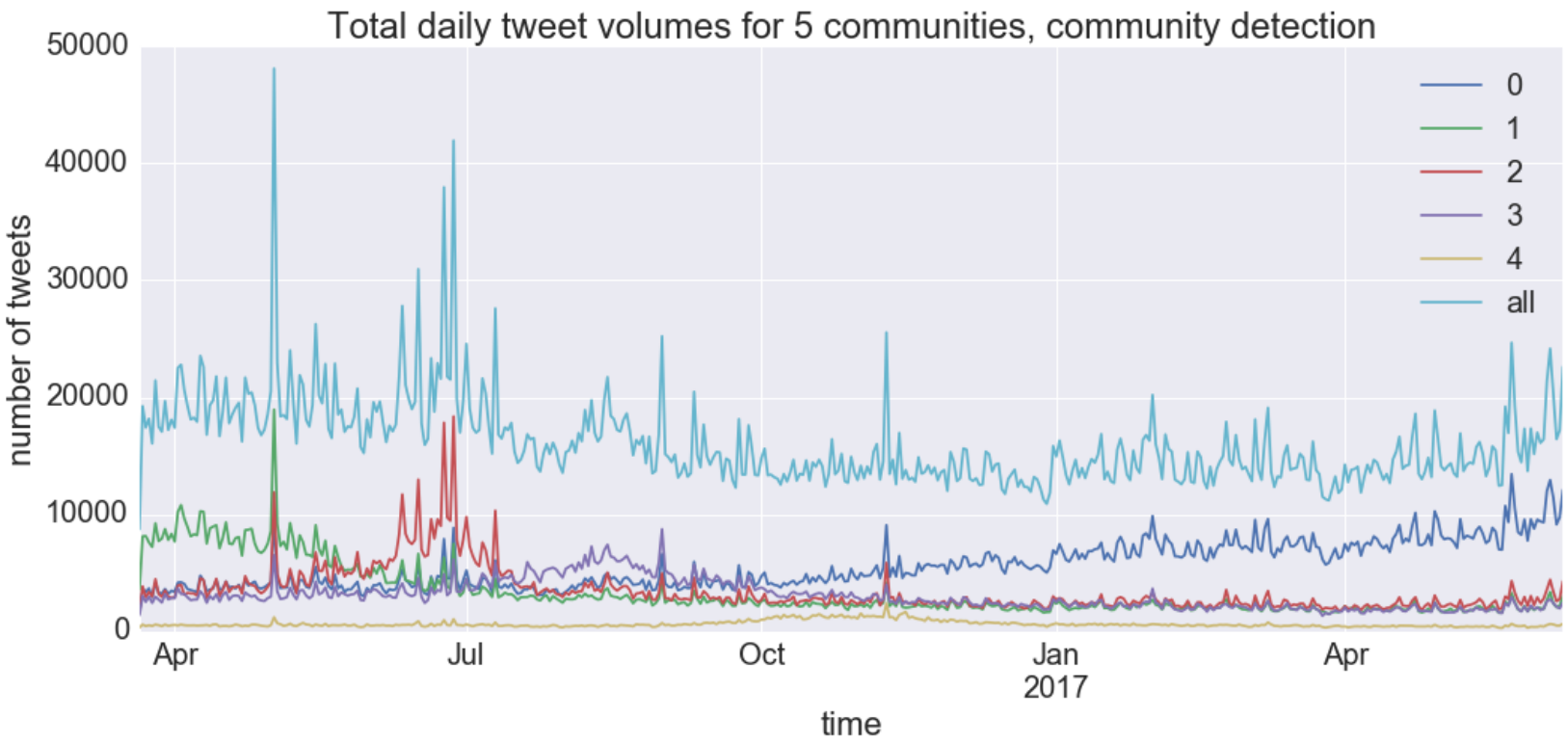


Figure 20: daily volumes for five communities retrieved by community detection.

6. Discussion

6.1 Assessment of the approach

The results demonstrate the possibility of identifying distinct communities within a population of tweeters based solely on their behavioural patterns over time. Moreover, the consistency in the patterns identified by both hierarchical clustering and community detection indicates that the effect is largely independent of the selected clustering method.

The method appears to work particularly well for behaviour over weekly intervals, and is accordingly well-suited to identifying communities who respond in a distinctive way to events that take place over the course of one or multiple weeks. The interpretation presented above proposes that a community of students can be identified from a surge in their activity over the summer, for instance, with likely Americans distinguished from Brits on the basis of their higher levels of activity in the run-up to US rather than UK political events.

The approach does have clear limitations, though. Perhaps the foremost of these relates to the time intervals for which it is effective. As observed above, identifying similar users based on correlations in their daily patterns of activity is made difficult by how little most users tweet (see 4.1.b(i)). But limiting the process to behaviour covering weekly intervals (or longer) means that communities can only be distinguished based on responses to certain sorts of events (typically those that last for at least one week).

Moreover, events across multiple domains happen concurrently. As discussed in 5.3.d(ii), it is difficult to distinguish between activity related to the EU referendum and Euro 2016 on a temporal basis, since both happen in the same period. Equally, it is hard to say to what extent the surge in activity noted for the community of proposed "student sport fans" over the summer of 2016 (see 5.3.d(iii)) is on account of school breaking up, and how much of it is a result of the quadrennial occurrence of the Olympics.

It should also be noted that not every aspect of life is based around major events. Many indicators of interest or cultural background may be communicated irregularly, and at times that do not relate specifically to what is being expressed. In this sense, evidence of interest in sport or politics – both of which revolve around major, newsworthy occurrences – will always likely be the easiest to glean from such an approach.

The nature of Twitter itself should also be considered. The irregularity with which most users tweet does not necessarily demonstrate infrequent use of Twitter, but might instead be seen to reinforce findings that the majority of activity on the platform consists of browsing (Khoo, 2014). Relying on tweet activity is accordingly limited for inferring community membership – some further assessment of the times at which people are online and other elements of their activity, such as liking the tweets of others, may give a more accurate picture of community membership based on temporal behaviour.

Similarly, it has been observed that Twitter users – and particularly those that tweet – do not necessarily reflect the demographic of the population as a whole (Ruths and Pfeffer, 2014). While there are indications that it is possible to work around demographic limitations (Wang et al., 2014), this is at least something to be aware of for any applications of the approach to questions of community detection and population integration in the "real" world.

6.2 Limitations of the study, and areas for future investigation

Beyond the broader limitations of the proposed approach, there is a range of areas in which this study and the assessment it provides of the approach is currently deficient.

While it is clear from our results that identifying communities based on representations of temporal behaviour is possible, more needs to be done to establish the significance of these communities. Our interpretations of the communities returned rely on inferring common interests or characteristics from a consideration of events that peaks in community behaviour seem to correspond to. A preliminary assessment of tweet content was carried out at the most prominent activity peaks to ascertain the events of interest (the most commonly used terms around the peaks were extracted and used for this), but more needs to be done. For establishing that one community is interested in football while another is more interested in politics, for instance, a simple binary classifier could be trained and run on all of the tweets for each community. If the proportions of football- and politics-related tweets differed as expected, we could assert the distinctness of these communities more strongly.

Following Campigotto and Guillaume (2014), reproducible communities are likely to be the most significant. Further analysis of the consistency of the communities returned across clustering methods would accordingly be beneficial. While preliminary analysis suggests that some 70% of users are placed in the same community (based on behavioural profile) by both the community detection and hierarchical clustering methods, much more work could be done here. Interestingly, better crossover was actually found to exist between community detection and *average* hierarchical clustering, rather than the complete method focused on above. Assessing consistency may also prove worthwhile since pruning communities down to their core members could result in more distinct and informative behavioural profiles.

The lack of perfect crossover between methods might also be treated as positive, with each putting a different slant on the data. Pulling five communities from average hierarchical clustering, for example, reveals possible latent communities that were not apparent in the other methods (see the average daily activity of communities 3 and 4 in Figure 21). Some way of combining these methods, perhaps by identifying core members of communities that appear across all of the approaches and allocating inconsistent members to less common behavioural profiles that they also fit, could prove the most effective. Further to

this, there is clearly also scope for the consideration of alternative clustering and community detection methods, beyond those looked at here.

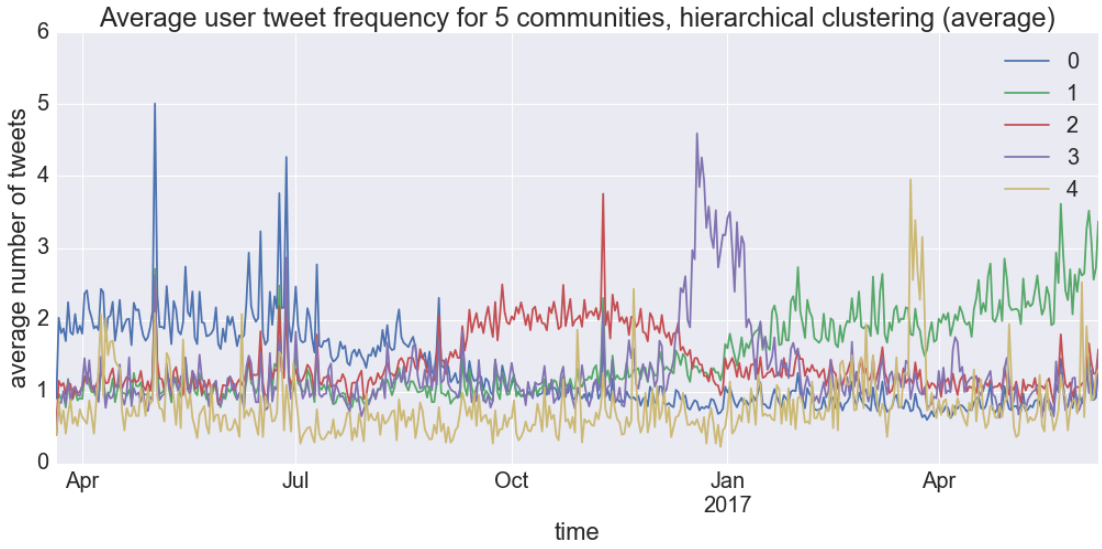


Figure 21: average daily tweet frequency for five communities retrieved by hierarchical clustering.

More experimentation is required to assess the effectiveness of the approach using different time intervals, too. While weekly intervals appear to produce results, a range of periods covering everything between a few days and two weeks would likely also be effective. Similarly, we consider activity over 14.5 months here, but both longer (with people who tweet less regularly) and shorter (with people who tweet more regularly) periods could be considered and may demonstrate different effects. Longer periods could also be divided, with clustering carried out at multiple points, either to assess consistency or find evidence of changes in community membership.

It may also be possible to overcome the problem of tweet sparsity to use shorter time periods: perhaps by removing users that tweet too infrequently for correlations to be found effectively, before placing them back into communities based on their similarity to users at lower levels of temporal granularity (e.g. over weekly periods). The potential use of categorical data may have been dismissed prematurely, too. Just as users that tweet too infrequently might be removed and placed back into communities later on, users who tweet too frequently – with 1's recorded for every period in categorical data – could be temporarily taken out of the population, and later placed in communities to which they seemed suited based on another level of analysis.

This hints at the scope for multi-level analysis. Communities identified based on weekly behaviour, for example, could be assessed further at progressively higher levels of temporal granularity. The initial community detection process would then constitute merely a preliminary filter, beyond which sub-groups in that community could be identified with ever-greater detail.

Further consideration of how far smaller groups can be meaningfully detected is also necessary more broadly, especially for applications to research on

integration. Many minority groups in the real world constitute 1% or less of a total population, but the communities considered here have rarely made up less than 10% (although much smaller communities were also detected). Moreover, the groups we have identified appear to relate primarily to shared interests, rather than shared religion, ethnicity, nationality, income, or political outlook, for instance – all of which are more commonly considered when assessing the integration of a population. Again, that might change with consideration of smaller communities.

Equally, this study only applies the approach to one population: experimenting with different populations, defined in different ways, is an important step towards establishing its wider applicability. It would be particularly interesting to determine whether the wealth of negative correlations between users found for this population is typical, or anomalous – although more work to establish the cause and significance of these negative values would be worthwhile either way. The typicality and significance of the apparently scale-free nature of the networks resulting from high-thresholded correlation matrices would also be good to consider.

Finally, following the behaviour of communities further back in time would help better establish the veracity of the profiles assigned to them. Some users would need to be (temporarily) removed for this, since their last 3,200 tweets may not go back far enough. For many, though, this should be possible. An obvious example for which going further back would be useful is the supposed community of students. Finding out whether or not this community exhibits the same spike in behaviour over multiple summers would strengthen or largely disprove the current theory.

6.3 Measuring integration: next steps

We now return to the original motive of the study, which was the measurement of integration in virtual space. We proposed that this could be useful either for bolstering existing measures of the integration of real-world populations by offering an assessment of their integration in the virtual world, or for determining the extent to which virtual populations, unbounded by geography, are integrated.

The approach outlined above already offers a means of intuitively assessing the integration of a population, or of the extent to which communities are integrated into a larger whole. By visualising the behaviour of populations and sub-groups over time, it becomes immediately apparent how far users "tweet together". In the case of the five communities returned by community detection and considered in detail above (see section 5.3.d), for instance, it is striking that in the build-up to the US Presidential Election all but one of the five is largely inactive, with the activity of at least three in decline. The activity of the fifth community, meanwhile, manifestly spikes, indicating that the primary interests or concerns of this community differ from those of the rest of the population,

with that community apparently less well integrated than the others. Another striking example is community 3 (purple) in Figure 21 on page 35.

However, this sort of visual assessment is difficult to quantify. An obvious area for further work is therefore into ways of building on the above to produce quantitative measures of integration. For example, the extent to which an identified community should be seen as integrated into a population could be measured by calculating the correlation between the patterns of behaviour of the community and larger population, as was done for users in this study (see 4.1.b). Coming from a different direction, Moskvina and Liu (2016) have developed a measure of the "togetherness" of two networks, and this could be built on for establishing how well integrated two communities are.

As a final thought, most of this work has concerned the process of decomposing a population into sub-communities, and some attempt to measure the difficulty of that decomposition might be a way forward. Before even starting the community detection process, an assessment of the user correlation matrices could be informative: here we had many negative values, which arguably indicate fairly low levels of integration (users not just tweeting at different times, but some tweeting less when others tweet more, and vice versa); another population might return large numbers of relatively high correlation values, likely indicating greater integration. The distances between communities returned by hierarchical clustering may also be significant, with a tightly packed dendrogram and little to differentiate between clusters another possible indicator of high levels of integration.

7. Conclusion

The approach to identifying sub-communities in populations of tweeters outlined here is a step towards the effective measurement of integration online for multiple reasons. First of all, considering tweeting patterns over time is a fast and simple alternative to looking at tweet content or explicit network relationships when clustering similar users. It covers a facet of Twitter activity – the times at which people tweet – that has been largely overlooked, yet seemingly allows for the the meaningful identification of communities within populations of tweeters. While this temporal approach can be effective alone, it could also be combined with more typical approaches for mutual benefit.

Secondly, considering behaviour over time enables the visual disentanglement of community contributions to population behaviour. Not only is inferring the likely makeup of communities possible, but times at which communities "tweet together" can be easily identified. This allows for an intuitive assessment of how integrated a population is as a whole, and of how far the behaviour of individual communities appears to be integrated into that of the larger population. Furthermore, where divergences are visible, it is possible to see the nature of the events that these divergences relate to. It then becomes possible to not only assess broad levels of integration, but the extent to which populations and communities are integrated in terms of their responses to certain sorts of topics.

Beyond the further work required to refine the approach, the next step is to make these intuitive assessments measurable. A number of possible future paths building on this work are proposed, including simple correlation measures between the tweet patterns of communities and populations, and a measure of how easy it is to identify communities within a population in the first place.

Word count: 11,561

Bibliography

Akcora, C. G., Bayir, M. A., Demirbas, M., & Ferhatosmanoglu, H. (2010). Identifying breakpoints in public opinion. *Proceedings of the First Workshop on Social Media Analytics - SOMA 10*. doi:10.1145/1964858.1964867

Asano, E. (2017). How much time do people spend on social media? http://www.socialmediatoday.com/marketing/how-much-time-do-people-spend-social-media-infographic [Accessed 2017-08-27]

Aynaud, T. (2009). Community Detection for NetworkX. http://python-louvain.readthedocs.io/ [Accessed 2017-07-20]

Bagheri, E., Du, W., Fani, H., Feng, Y., Zarrinkalam, F., & Zhao, X. (2015). Temporal Identification of Latent Communities on Twitter. *CoRR, abs/1509.04227*.

Bie, T. D., Lijffijt, J., Mesnage, C., & Santos-Rodriguez, R. (2016). Detecting trends in twitter time series. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. doi:10.1109/mlsp.2016.7738815

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,2008*(10). doi:10.1088/1742-5468/2008/10/p10008

Bodine-Baron, E., Helmus, T., Magnuson, M., & Winkelman, Z. (2016). Examining ISIS Support and Opposition Networks on Twitter. doi:10.7249/rr1328

Burger, J. D., Henderson, J., Kim, G. and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, pp.1301-1309.

Campigotto, R., & Guillaume, J. (2014). The Power of Consensus: Random Graphs Still Have No Communities. *Lecture Notes in Social Networks Social Network Analysis - Community Detection and Evolution,*145-164. doi:10.1007/978-3-319-12188-8_7

Casey, L. (2016). The Casey Review: a review into opportunity and integration. UK Government publication, available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/575973/The_Casey_Review_Report.pdf [Accessed: 2017-07-10]

Cihon, P., & Yasseri, T. (2016). A Biased Review of Biases in Twitter Studies on Political Collective Action. *Frontiers in Physics,4*. doi:10.3389/fphy.2016.00034

Cross Validated (2015). Using correlation as distance metric (for hierarchical clustering). https://stats.stackexchange.com/questions/165194/using-correlation-as-distance-metric-for-hierarchical-clustering [Accessed 2017-08-19]

DCLG (Department for Communities and Local Government) (2012). The Citizenship Survey. https://discover.ukdataservice.ac.uk/series/?sn=200007 [Accessed 2017-08-26]

Durkheim, E. (2013)[1893]. *The division of labour in society*. Houndmills: Palgrave Macmillan.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer,66*(4), 568-578. doi:10.1080/00330124.2014.907699

Grimmett, G. (1999). *Percolation*. Second edition, Springer-Verlag: New York.

Hagberg, A., Schult, D. and Swart, P. (2008) "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), Pasadena, CA USA pp. 11–15

Heath, A., Fisher, S.D., Rosenblatt, G., Sanders, D., & Sobolewska, M. (2013) *The Political Integration of Ethnic Minorities in Britain.* Oxford University Press: Oxford

Himelboim, I., Mccreery, S., & Smith, M. (2013). Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication,18*(2), 40-60. doi:10.1111/jcc4.12001

Hubspot (2009). State of the Twittersphere. http://cdn2.hubspot.net/hub/53/blog/sotwitter09.pdf [Accessed 2017-08-27]

Jones E, Oliphant E, Peterson P, et al. (2001). SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/ [Accessed 2017-08-29].

Khoo, C.S.G. (2014). Issues in information behaviour on social media. *LIBRES, 24*(2), 75-96.

Krasodomski-Jones, A. (2016). Talking to Ourselves? Political debate online and the echo chamber effect. Demos report, available at https://www.demos.co.uk/wp-content/uploads/2017/02/Echo-Chambers-final-version.pdf [Accessed 2017-07-05]

Margetts, H., John, P., Hale, S., and Yasseri, T. (2015) *Political Turbulence: How Social Media Shape Collective Action.* Princeton, NJ: Princeton University Press.

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Presented at PyHPC2011: https://www.scribd.com/document/71048089/pandas-a-Foundational-Python-Library-for-Data-Analysis-and-Statistics [Accessed 2017-08-28]

McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–445.

Moskvina, A., & Liu, J. (2016). Togetherness: An algorithmic approach to network integration. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. doi:10.1109/asonam.2016.7752239

Mugan, J., Mcdermid, E., Mcgrew, A., & Hitt, L. (2013). Identifying groups of interest through temporal analysis and event response monitoring. *2013 IEEE International Conference on Intelligence and Security Informatics*. doi:10.1109/isi.2013.6578816

ONS (Office for National Statistics) (2012). 2011 Census: Ethnic group, local authorities in England and Wales. http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/rft-table-ks202ew.xls [Accessed 2017-08-20]

Palla, G., Derényi, I., & Vicsek, T. (2006). The Critical Point of k-Clique Percolation in the Erdős–Rényi Graph. *Journal of Statistical Physics,128*(1-2), 219-227. doi:10.1007/s10955-006-9184-x

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying User Income through Language, Behaviour and Affect in Social Media. *Plos One,10*(9). doi:10.1371/journal.pone.0138717

Rao, D., Yarowsky, D., Shreevats, A., Gupta, M. (2010). Classifying Latent User Attributes in Twitter. *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*. SMUC pp. 37–44.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science* 346:1063–4. doi: 10.1126/science.346.6213.1063

Statista (2017). Number of monthly active Twitter users worldwide. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [Accessed 2017-08-27]

Twitter (2013). New Tweets per second record, and how! https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html [Accessed 2017-08-27]

Varol, O., Ferrara, E., Davis, C., Menczer, F. and Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. arXiv:1703.03107

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting,31*(3), 980-991. doi:10.1016/j.ijforecast.2014.06.001