Sustainability Research Institute
SCHOOL OF EARTH AND ENVIRONMENT

UNIVERSITY OF LEEDS

Using Twitter data to identify networks of interest in minority policy topics

Malcolm Morgan, Gavin Killip, Marina Diakonova

April, 2019

No. 118

SRI PAPERS

SRI Papers (Online) ISSN 1753-1330

First published in 2019 by the Sustainability Research Institute (SRI)

Sustainability Research Institute (SRI), School of Earth and Environment,
The University of Leeds, Leeds, LS2 9JT, United Kingdom

Tel: +44 (0)113 3436461
Fax: +44 (0)113 3436716

Email: SRI-papers@see.leeds.ac.uk
Web-site: http://www.see.leeds.ac.uk/sri

About the Sustainability Research Institute

The SRI is a dedicated team of over 20 researchers working on different aspects of sustainability. Adapting to environmental change and governance for sustainability are the Institute's overarching themes. SRI research explores these in interdisciplinary ways, drawing on geography, ecology, sociology, politics, planning, economics and management. Our specialist areas are: sustainable development and environmental change; environmental policy, planning and governance; ecological and environmental economics; business, environment and corporate responsibility; sustainable production and consumption.

Disclaimer

The opinions presented are those of the author(s) and should not be regarded as the views of SRI or The University of Leeds.

Using Twitter data to identify networks of interest in minority policy topics

© [Malcolm Morgan 2018]

Email: m.morgan1@leeds.ac.uk

Contents

Abstract

This paper presents a method for using Twitter to ascertain insights into the social structures of a minority policy topic. The example of the UK domestic building retrofit sector is used, as an example of a sector with limited success in achieving policy goals.

The working paper outlines how it is possible to gather data from Twitter on a specific sector of interest and analyse this data to find distinct clusters and structures within the chosen sector. Data was gathered for about 17,000 accounts including the friends' relationships between accounts and 36 million tweets produced by these accounts. The paper presents some of the challenges and limitations of using Twitter data as well as opportunities for the method to be used in other sectors.

Keywords: Retrofit, Twitter, Social Network Analysis, Housing

Submission date [23-10-2018]; Publication date [01-04-2019]

About the Authors

Malcolm Morgan

Malcolm works for the Sustainable Research Institute and the Institute for Transport Studies at the University of Leeds. He is a Civil Engineer with an interest in low-carbon infrastructure, particularly in the housing and transport sectors. Malcolm's research focuses on how big data techniques can be applied to understand better our current and future infrastructure needs and how what changes are required to transition to a low-carbon economy.

Gavin Killip

Gavin is an inter-disciplinary researcher interested in finding solutions for a more sustainable built environment. He takes a broad 'socio-technical systems' approach to investigate how technology and behaviour evolve and affect each other, with the ultimate goal of proposing a positive change by understanding better the workings of complex systems. Gavin's main research focus has been on existing housing in the context of climate change mitigation - investigating how 2050 climate change targets might be met through the markets for property (sales and rentals) and repair, maintenance and improvement. He is also active in the field of research on multiple benefits of energy efficiency, and on more general issues of governance for a low-carbon transition.

Marina Diakonova

Marina Diakonova is a data scientist based in Oxford University's Environmental Change Institute. She is part of the METER study, which aims to assess the potential of Demand Side Response by matching household electricity consumption to the underlying human activities. Marina got her PhD in the Warwick Complexity DTC, after which she worked on multilayer networks with Maxi San Miguel and Vito Latora. She is interested in using data science and creative visualisation to advance interdisciplinary projects tackling a range of social questions.

1 Introduction

This paper sets out a method for using Twitter data to identify a network of interest in a minority policy topic. In this case, the minority policy topic is 'retrofit' - the renovation of existing homes in order to significantly improve their energy efficiency and reduce the associated emissions of carbon dioxide. The method was developed for an inter-disciplinary research project on retrofit in the UK ('Governance of Low-carbon Innovations for Domestic Energy Retrofits' or GLIDER), but the approach may be of interest for other topics where a community of interest exists. Indeed, it seems probable that there are many such topics across political and policy-related discourse, where goals, targets, and objectives have been identified, but where implementation at any sufficient or meaningful scale still lags behind. Such topics of discourse may be promoted and sustained by advocates, pioneers, policy-makers and others. Analysing such communities of interest may be of interest to researchers of political science, policy-making; and an investigation of the network of interest may reveal findings that network members themselves find useful and insightful.

The method is unlikely to apply universally to all topics of minority policy interest, however, not least because the source data from Twitter is not representative of the whole population; Twitter users in the UK are rather skewed towards younger male users from professional and managerial work backgrounds (Sloan et al., 2015). This means the social structures and topics of interest among Twitter users cannot be assumed to represent the wider offline community. Even so, we believe that the methods described here may be of wider relevance than just the community of interest around retrofit. In particular, the low cost of gathering and analysing Twitter networks could provide an easy way to make a first approximation of social structures in under-studied groups.

The aim of this paper is primarily to describe the methods used, and not to give a detailed analysis of results. Where examples are given to illustrate the methods, they are taken from our work on retrofit, but this paper does not set out to analyse or reflect on the significance of the work for debates on retrofit.

2 Literature review and background

This section will introduce the concept of 'a topic of minority policy interest' and why retrofit meets that definition. It will then introduce the field of computational social science with relevance to Twitter and highlight the specific characteristics of Twitter data.

2.1 Retrofit as a topic of minority policy interest

What is a topic of minority policy interest? We define the term in relation to several components:

- The existence of one or more policy documents identifying the topic as being important to the achievement of future goals;
- A gap between the level of activity implicit in the policy statements and the level of activity observed in reality;
- At least one identifiable group of people who are active in campaigning or communicating on the topic in order to reduce the gap between policy rhetoric and implementation.

It is, therefore, less about the absolute level of interest, but much more about the gap between policy rhetoric and reality. Topics of minority policy interest may be quite diverse, including topics based on geography (neighbourhoods, towns, cities, or regions wishing to be more widely promoted); on research evidence (e.g. climate change). However, it does not include all topics of minority interest, as it explicitly excludes those where there is not at least one policy document in support of the broad goals of a topic's supporters.

Housing retrofit fulfils all three of these criteria that we have proposed for the definition of a topic of minority policy interest:

- Numerous policy documents exist at international level and national level (HMSO, 2008; IPCC, 2014; International Energy Agency, 2017; CCC, 2018).
- Modelling studies and policy documents acknowledge the gap between the level of current activity and what is needed in future (Skea et al., 2009; BEIS, 2017).
- Groups involved in the promotion of retrofit include National Energy Foundation, Passive House Institute, Parliamentary Renewable and Sustainable Energy Group – among many others.

Retrofit is significant among energy policy-makers and researchers, as well as in parts of the construction industry and among community groups concerned to promote solutions to climate change. High-level strategic documents regularly cite the importance of retrofitting entire national building stocks if energy and climate targets are to be met – at international, national, and regional levels (IPCC, 2014; Patrick et al., 2014; BEIS, 2017). However, the rhetoric of retrofit 'roll-out' does not translate straightforwardly into action on the ground, despite the availability of mature and cost-effective technologies (Stafford et al., 2011). One set of reasons for this is the fragmented nature of the market for housing repair and maintenance; the fact that the challenge exists at multiple scales simultaneously; and the need for novel multi-sector networks of stakeholders to achieve policy goals. There is no right answer to such 'messy' and 'wicked' problems (Rayner, 2010).

These considerations highlight the importance to the policy process of understanding social structures, management, and organisation. For retrofit, these topics are revealed as important in case studies, where meticulous attention to detail and good communication among members of the project team are key to success. There have been calls for the creation of a new 'integrator' or 'coordinator' role to manage the often disparate and fragmented workforce and achieve greater quality assurance (WBCSD, 2009; Killip, 2013). Looking beyond the level of individual buildings and projects, there is a growing recognition in policy and industry debates that delivery of retrofit at any significant scale may require social innovations, such as new business models (Killip et al., 2014; Mlecnik et al., 2018). The shift of focus can be thought of as moving away from the techno-economic questions of 'what', 'how much' and 'when' towards the social and managerial questions of 'how' and 'by whom' (Janda and Parag, 2013).

Retrofit represents a different kind of activity and a different group of stakeholders from what has gone before in residential energy policy (Owen et al., 2014). This change, should it take place at any significant scale in the future, has profound implications for jobs, tasks and responsibilities, i.e. for the social organisation of work.

By investigating the social network of debate around retrofit, we seek to identify the social structures that could inform policy in the broadest sense. Who are the stakeholders engaged in this topic? What are the patterns of communication among the network? Can any network-level features, such as clusters (communities with close reciprocal links) and structural holes, be identified? Can any weak ties be identified, which could be important bridges from one community to another, following the theory of weak ties (Granovetter et al., 1973)? What lessons, if any, can be inferred for future interventions from policy-makers or other actors? Can an investigation of retrofit as an expression of network structures lead to new and useful insights – for retrofit, but also for other policy debates?

This working paper will focus on the methods of gathering data on a minority interest Twitter community and show that distinct sub-communities exist. Follow on publications will then address the interpretation of these sub-communities and their policy implications.

2.2 Twitter

Twitter is a social media platform that allows users to publish short public messages (tweets) on any topic to their personal timeline. Users can 'follow' other users to receive their tweets on an ongoing basis. Users can also 'like' other users' tweets or 'retweet', i.e. repost another's tweets to their timeline. Finally, users can mention another user within their tweet by using the @ prefix to their username. These four actions (Following, Liking, Retweeting, and Mentioning) are the core of Twitter's social network. Twitter also makes use of 'hashtags', whereby the prefix # to any string of characters is a means of sign-posting users to other tweets on the same topic. Twitter activity is in the public domain (unlike other social media platforms like Facebook, Instagram) and is accessible via the Twitter website, as well as via the Twitter app. The Application Programming Interface (API) for Twitter allows data from the platform to be gathered, although API rules create practical and technical constraints (which are discussed below).

2.3 Structure of Interactions on Twitter

The four actions mention above, produce distinct structures within the social network. Each of these actions is described below. Each action is directional in nature, although the direction is not always clear. The implications of directionality in twitter are discussed further in sections 2.4 and 2.8.3.

2.3.1 Following

On Twitter, one user can follow another user, as shown in Figure 1. Unlike other social networks, following is unidirectional and thus need not be reciprocated (B follows C, but C does not follow B). Within the Twitter API, the distinction is made between Followers (accounts that follow you) and Friends (accounts that you follow).

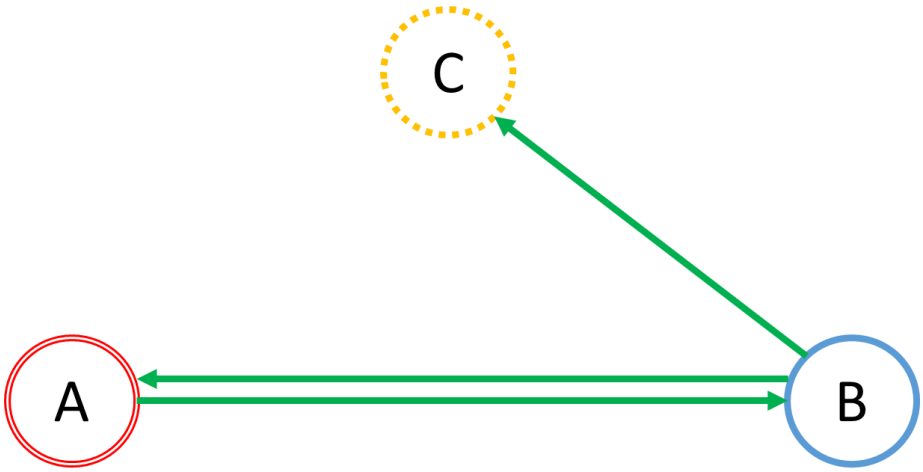

Figure 1: Simple example of following on Twitter, Arrow signifies the direction of the connection.

Within this simple example, A has one 'friend' (B) and one follower (B). B has two friends (A and C) and one follower (A); finally, C has no friends and one follower (B).

2.3.2 Tweeting and Retweeting

A users' timeline consists of any tweets they have posted. This can include both original tweets, retweets of others, or retweets of the user's earlier tweets. Figure 2 demonstrates this concept showing three accounts each with three tweets in their timeline. The colour of the tweet signifies the original author of the tweet. Due to the way Twitter records retweets, it is not possible to track a retweet across multiple users. For example, in Figure 2 it is not possible to determine if user C retweeted user A directly or retweeted user B's retweet of A, as represented by the dashed line. Some researchers have circumvented this problem by assuming that retweets only travel along the followers' network and thus provides a method for identifying the A-B-C path of retweeting (Kwak et al., 2010). Some cross-validation of this method may be possible by examining the timing of retweets.

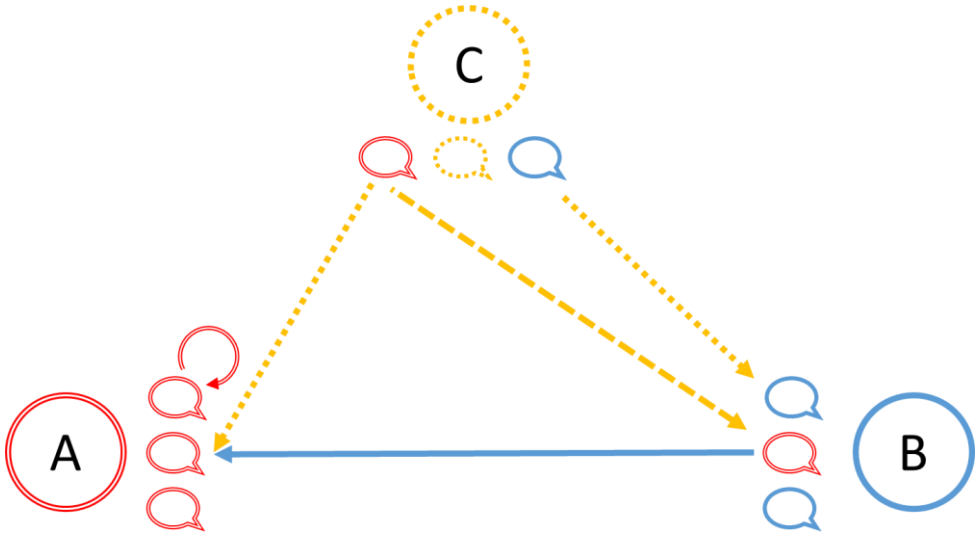

Figure 2: Simple example of tweeting and retweeting. Each account has three tweets in its timeline, with the colour of the tweet representing the original author. Arrow signifies the direction of the connection.

2.3.3 Like (Favourites)

Users can mark tweets by themselves or others as a “favourite” tweet. Confusingly, within the Twitter user interface this is referred to as “like” but referred to as “favourite” within the Twitter API. Users can, upon reading it, “like” any tweet by themselves or others. It is possible to like a retweet although in practice this appears to be uncommon, accounting for 0.0001% of favourites in our sample.

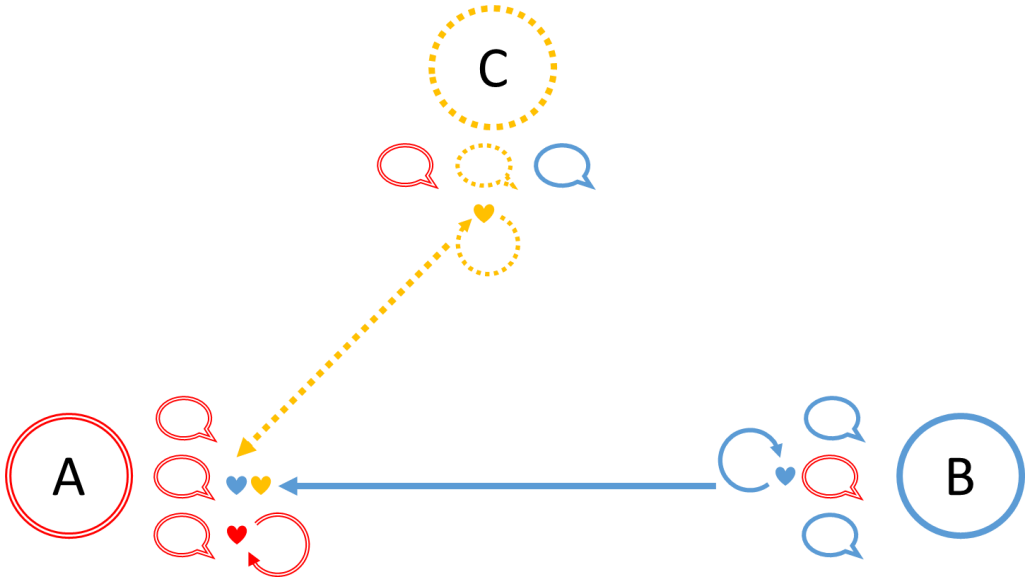

Figure 3: Simplified example of favourite tweets across multiple accounts, Arrow signifies the direction of the connection.

2.3.4 Mentions

Mentions cannot be gathered directly from the Twitter API, but instead can be derived from the contents of Tweets by looking for “@account.” A tweet can contain any number of mentions, constrained only by the character limit of the tweet, at the time of the data collections this was 140 characters.

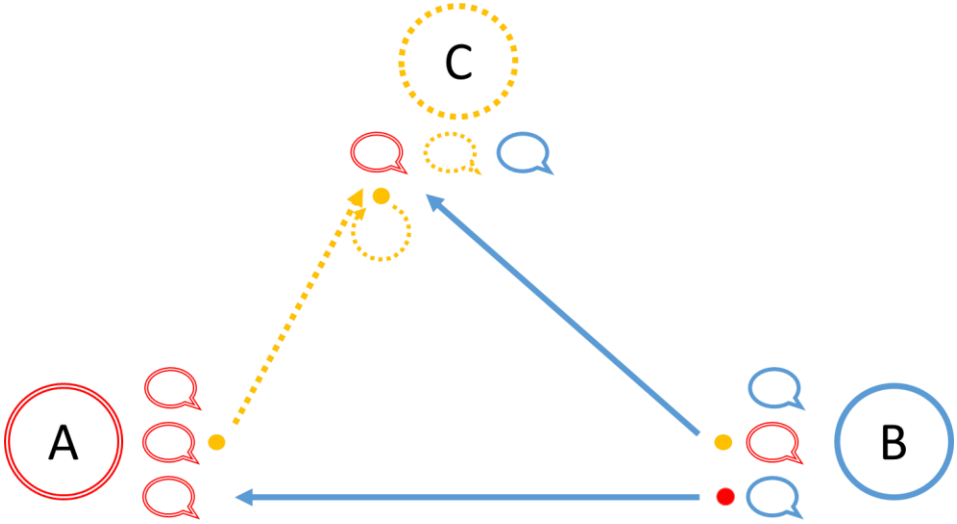

Figure 4: Example structure of mentions. Arrow signifies the direction of the connection.

2.4 Computational Social Science

Much of the research on Twitter is interested in Twitter behaviour in its own right, not necessarily using Twitter as a proxy for off-line social structures. While computational analysis is growing in its sophistication, it is hard to find any attempts to draw conclusions from online activity for the off-line context in the real world in the academic literature. This reflects the immaturity of the field and the lack of consensus on the underlying meaning of Twitter data. Taha Yasseri's paper (Cihon and Yasseri, 2016) argues that the Twitter research literature is only unified in the source of its data; "most often authors do not consider the expansive political and social theoretical literature in their analyses of online social phenomena. Instead, they provide case studies and methodological developments exclusively for Twitter research … we find that many papers fail to support their choice of methodology within the greater literature."

Despite the immaturity of the field, some clear findings are now emerging. First, not all people use Twitter. For example, Sloan et al (2015) analysed professional class and age on UK's Twitter in relation to census data for the whole population. He suggested that some socio-demographic groups like the managerial-professional are almost perfectly represented on Twitter, while others, like the technical ones, are under-represented. Perhaps unsurprisingly, there is an overrepresentation of young people on Twitter, (Sloan et al., 2015).

Secondly, the way Twitter is used is not uniform. The type of use falls somewhere between a social network and news media (Kwak et al., 2010). Valenzuela et al (2018) argue that Twitter is based on networks of relatively weaker social ties (e.g. acquaintances, work contacts), compared with the stronger ties of friendship and kinship on Facebook, making Twitter a good vehicle for 'injecting novel information' (Valenzuela et al., 2018). Grabowicz suggests that these different uses of Twitter can be distinguished by the method of communication. A retweet takes less effort and represents and an "information diffusion event", where a user is simply relaying information across their network. Whereas mentions are personal messages, and thus their number can be a proxy for the strength of the corresponding social tie (Grabowicz et al., 2012).

Finally, methodological choices relating to the gathering and analysis of Twitter data can have a significant impact on the results and must be grounded in theory (Cihon and Yasseri, 2016). Three common source of bias identified are, researcher introduced bias due to the sampling method for selecting the tweets and/or accounts to be studied, the 'black box' of the Twitter algorithms which can restrict the researcher control over sampling, and innate biases in the Twitter community.

Overall, this nascent field has yet to establish a strong theoretical background for the interpretation of Twitter networks. The rest of this section will cover the recent methodological developments in the field.

2.5 Data Collection Methods

Twitter provides multiple ways to access its data and therefore research methodologies for data collection vary based on the source of data. The Search API (the focus of this paper) provides the ability to retrieve historical data, with certain limitations (See section 2.8). Alternatively, the

Streaming API provides a real-time feed of current activity on Twitter. The method of data collection significantly affects the type of results gathered. Some of the major differences, from a research perspective, for each method, are listed below:

Search API

- Data pulled from Twitter based on search criteria;
- Allows research of historical interactions, with limitations (See section 2.8.2);
- Allows collection of friends/follower networks (See section 2.3);
- Requires targeting of specific, named accounts or keywords;
- Sampling process controlled by the researcher;
- Restrictions on the speed of data collection (See section 2.8.1).

Streaming API

- Data pushed from Twitter in real time;
- Provides a variable sample (1% - 40%) of all tweets (Bright Planet, 2018);
- Data collection must begin before the period of study;
- Sampling process controlled by Twitter;
- Friends/followers networks not provided;
- No overall limits on the volume of data collected.

A further distinction can be made between researchers who used freely available APIs and those who pay or obtain special permission to access data in greater volumes. These sample descriptions are further complicated by Twitter changing policies on data sharing (González-Bailón et al., 2014). As Twitter does not permit the bulk publication of tweets gathered from the API and the time of data collection affects both methods, it is impossible to reproduce the data collection process of any published Twitter research.

For researchers, the choice of API is a choice of sampling method. As sampling methods should reflect the research question, researchers have used a wide range of sampling methods: from considering all of Twitter (Kwak et al., 2010; Myers et al., 2014), all tweets in a given period (Grabowicz et al., 2012); a significant sample of all tweets (Bliss et al., 2012; Bild et al., 2015); to snowballing techniques based on a starting list (Beguerisse-Díaz, 2013).

This paper adopts a snowballing technique, which is appropriate as the intention is to study a specific sub-community which is conceptionally well defined, but whose membership is unknown prior to the study. A

2.6 Structure of Twitter Data

Full documentation of the Twitter API is beyond the remit of this paper and is available on the Twitter website. For the purposes of this paper, only five types of API request will be considered:

- GET users/show
- GET friends/list
- GET followers/list
- GET statuses/user_timeline
- GET favorites/list

Each of these API requests requires a Twitter account name as input and returns multiple results, as will be elaborated in the following sections.

2.6.1 Users

The “GET users/show” request returns basic details about a Twitter account when provided with either the account name (e.g. @RIBA) or the account identification number.

Table 1: Summary of data returned from the users request

Field Name	Description	Used in this analysis
description	User provided account description	Yes
statusesCount	Total number of tweets	Yes
followersCount	Total number of followers	Yes
favoritesCount	Total number of favourites/likes	Yes
friendsCount	Total number of friends	Yes
url	User provided web link	No
name	User provided full-text name	Yes
created	Date and time the account was created	No
protected	Is the account protected (it is not possible to recover data from protected accounts)	Yes
verified	Is the account verified	No
screenName	Twitter username without the @ prefix	Yes
location	User provided location	No
lang	Account language	No
id	Unique user account number	No
listedCount	Number of lists the account appears in	No
followRequestSent	Has a follow request been sent	No
profileImageUrl	Link to the user profile picture	No

2.6.2 Friends and Followers

The “GET friends/list” and “GET followers/list” requests return a list of all the friends or followers of a specified account. The structure of the data is identical to the “GET users/show” request, and so the table is not repeated.

2.6.3 Tweets

The “GET statuses/user_timeline” requests the users most recent tweets (including retweets).

Table 2: Summary of data returned from the tweets’ request

Field Name	Description	Used in this analysis
text	The text of the tweet	Yes
favorited	Is the tweet one of our favourites?	No
favoriteCount	How many people have favourites this tweet	No
replyToSN	Account name that this tweet replies to (if applicable)	No
created	Date and time created	No
truncated	Has the tweet been truncated	Yes
replyToSID	Tweet id that this tweet replies to (if applicable)	No
id	The unique id of the tweet	No
replyToUID	Account id that this tweet replies to (if applicable)	No
statusSource	Software or method used to create the tweet	No
screenName	Account name of the tweet’s creator	Yes
retweetCount	Number of times the tweet has been retweeted	No
isRetweet	Is this tweet a retweet	Yes
retweeted	Have we retweeted this tweet	No

2.6.4 Favourites / Likes

The “GET favorites/list” request returns the favourite tweets of a given account. The data structure is identical to that of tweets request, and so the table is not repeated.

2.7 Community detection on Twitter data

Community detection is a key analytical process for Twitter researchers. Using network topology and/or node (user or tweet) content, researchers can cluster similar nodes and provide insight into social systems on a macroscopic scale. There are a variety of techniques to achieve this, each with its own set of strengths and weaknesses. Weng uses the Infomap algorithm and tests the robustness of their results by applying a second community detection technique called Link Clustering (Weng, 2014). Conover et al. use a combination of two techniques, Raghavan's label propagation method seeded with node labels from Newman’s leading eigenvector modularity maximization (Conover et al., 2011). Many papers use one or more of these methods without demonstrating that the methods and definitions they are using are well justified for their specific problem.

Beguerisse-Diaz (2014) stresses the importance of the directionality of connections between accounts. Directionality can be:

- Simply removed;
- Removed by considering only bi-direction connections;
- Taken care of by weighing the bi-directed edges doubly.

They find that community detection is sometimes but not always affected by accounting for directionality. They use Infomap and Markov Stability (small N. For large N they reference Lambiotte's extension to the algorithm, but it is unclear whether its realisation is easy and/or feasible), developing an extension of Markov Stability to work on directed networks. Infomap, the authors say, leads to an over-partitioned description (interestingly, the same is suggested in (Yang et al., 2016)), and an unbalanced partition.

2.8 Practical issues and constraints

Access to the Twitter API is easy and free but has several limitations and constraints. The main three are rate limitation, volume limitation, and directionality.

2.8.1 Rate limitation

The Search API is rate limited, therefore only a limited number of requests for data can be made in any 15-minute period. As requests for distinct types of data have different rate limits, speed improvements can be gained by using the waiting time of one request to make a different type of request. It was found that on a fast internet connection the computer would spend around three minutes collecting data followed by twelve minutes waiting for the rate limit to reset. The software was designed to work on batches of 50 accounts at a time and gathered a complete set of data for each batch in around 75 minutes.

Rate limitation places practical restrictions on the maximum size of the network that can be studied. This requires a careful selection of accounts to prevent wasted time during data collection. It should be noted that the limit of the Twitter API is set on the number of requests, not the number of accounts. This distinction is important as data is provided in batches, thus multiple requests may be required for each account. Thus, accounts per minute could be increased by reducing the amount of data requested for each account. It is unclear how Twitter orders the results returned from the API, therefore it is not possible to ascertain if this would bias the results.

2.8.2 Volume limitation

Requests to the Twitter Search API will only return results up to a defined maximum:

- Tweets: 3,200
- Likes: 3,300

When the limit is exceeded, the most recent data is returned. This limitation will bias analysis against accounts that have a high number of tweets or have had an account for many years, as their tweets and likes will be underrepresented.

Although not physically limited, the number of friends and followers can reach practical limits associated with the rate limits discussed in the previous section. This is especially problematic

for followers, as popular accounts can have millions of followers and take many hours to collect data from.

2.8.3 Directionality

Twitter connections are directional between accounts, which is reflected in the network. However, it is worth reflecting on the meaning of this directionality, especially in the context of keyword analysis. For example, if account A retweets account B, it is taken as a link from A to B, yet the content of the tweet originated with account B. Conversely, a mention of account B within a tweet from account A contains content from account A. Table 3 summarises how different types of connections are represented in the network.

Table 3: Summary of link types within the network

Content Type	Description	Text originally from	Link Direction
Friend	A is following B	-	A to B
Likes	A likes a tweet by B	B	A to B
Mentions	A mentions B in a tweet	A	A to B
Retweets	A retweets a tweet by B	B	A to B

3 Methodology

This section outlines the four main stages of the method used in this research project. First, identifying a starting core of the network; second, collecting data from Twitter; third, constructing a network from the collected data; and finally, trimming the network down to a manageable size by removing less relevant accounts and connections.

Twitter provides an Application Programming Interface (API), which allows automated access to Twitter data (Twitter, 2018). The Twitter API can be accessed with a wide range of software, in this case, the R package *twitteR* was used (Gentry, 2016). The *twitteR* package provides convenient functions for connecting to the Twitter API and returning data.

3.1 Identifying the core of the network

The first stage of data gathering was to produce a list of organisations that were thought to be relevant to the UK retrofit sector. It was necessary to define a starting point for data collection, as the majority of Twitter users are unlikely to have any relevance to the UK retrofit sector. We postulated that by starting with a well-selected list of 'core' accounts and then radiating out from these accounts, it would be possible to gather most of the relevant accounts while minimising the number of non-relevant accounts gathered.

This was done by gathering expert opinion to produce a list of 56 organisations of importance to the UK retrofit sector, listed in Appendix 1. While this process introduces an element of subjectivity into the method, it is necessary and justifiable. The nature of the Twitter Search API requires at least one starting account to be specified. Using an expert-selected account is more likely to yield useful results than a randomly selected account. Furthermore, the use of multiple accounts in the

core reduces the risk of any one account biasing the results. Finally, this method would not be an unusual approach to selecting candidates for interview in a traditional snowballing technique.

Of these 56 organisations, 47 (84%) had a twitter account, with some organisations having multiple accounts. These accounts formed the initial core of the social network, although the assumption was that this list would be incomplete and biased, and thus only a starting point for enquiry.

3.2 Data Collection

As mentioned above the R package *twitteR* was used to access the Twitter API. A simple solution would be to use repeatedly the provided functions to gather the necessary data for each account. Unfortunately, the basic *twitteR* functions fail if the rate limit has been exceeded or the request is invalid. To allow the bulk collection of data, new functions were written. These new functions halt the generation of requests and wait if the rate limit is going to be exceeded, and can handle other common errors. The functions are listed in Section 7.1 Further enhancements were achieved by making multiple requests in parallel, thus exploiting the independent rate limiting on different types of API request.

Ultimately, by automating error and rate limit handling, and minimising the time spent waiting, the new functions maximise the number of Twitter accounts from which data can be gathered within a fixed period while reducing the time spent by the user to practically zero.

A function was written which took a list of twitter account names as input and then returned data for each of these accounts. The following data was gathered for each user account:

- Account Details: Summary metadata about users such as account description, number of tweets, location;
- Friends: The list of accounts and associated metadata a given user is following;
- Tweets: The tweets and associated metadata posted by a given user, up to 3,200 of the most recent tweets per account;
- Likes: The tweets and associated metadata that a given user has liked, up to 3,300 of the most recent likes per account.

As this method only finds connections in one direction, a network can only be constructed by snowballing out from an initial account, or accounts. Figure 5 demonstrates the snowballing technique for a simplified network. The first round (red) gathers connections out from account A and the second round (green) gathers accounts out from account B. In this case, all the connections between A and B are known but other connections are only known in one direction. Further rounds of data collection would be required to complete the network. Notice that the link between D and A is impossible to gather as there is no inward connection to account D. While this is a deficiency of the method, it is unlikely to affect significantly the results when collecting data in bulk as an alternative path to D would likely exist. Furthermore, an account that is not friended, retweeted, or liked by any other account in the network is unlikely to play a significant role in that network.

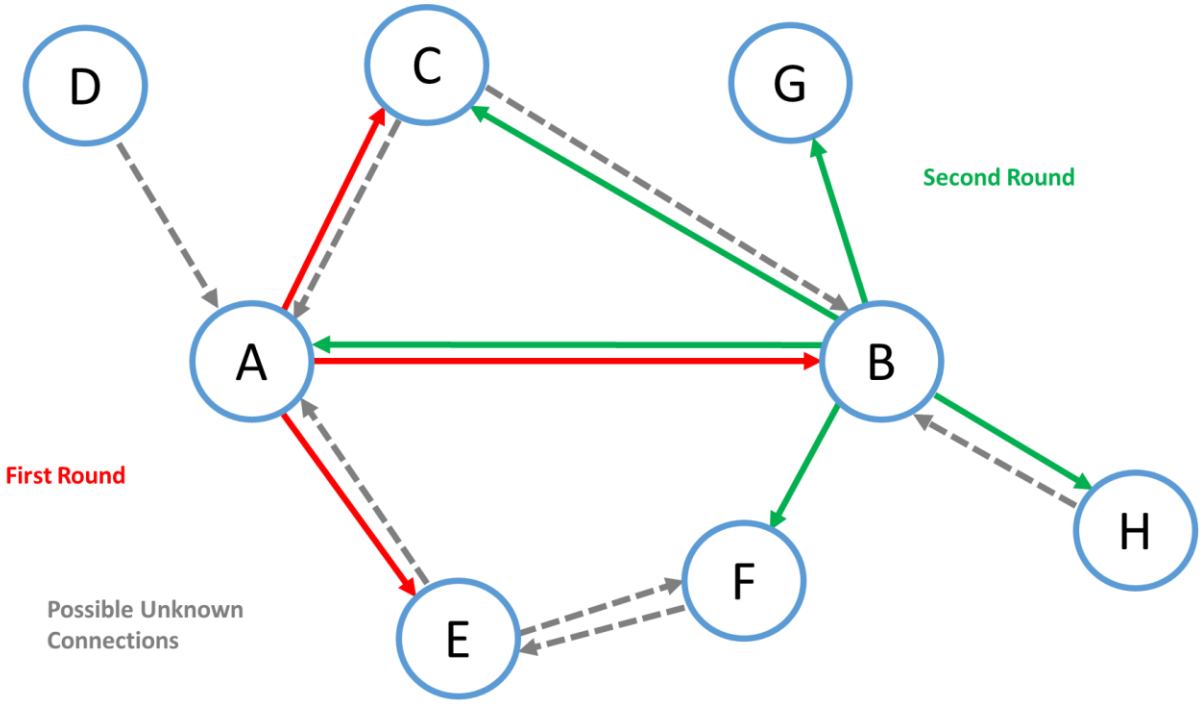

Figure 5: Snowball data collection process, highlighted known and unknown connections

If this method is extended to collect multiple accounts per round, then all the connections between the initial accounts are included. For example, if the first round of data collection covered A, B and C then all of the connecting between A, B, and C are known after the first round.

Data collection took place between 5th and 27th of November 2017 in three rounds as shown in Figure 6. It is not expected that the time of data collection would significantly bias results as accounts activity both during and before the data collection period was being gathered. However, a slight imbalance will exist between the accounts collected at the start and end of the data gathering period, as later accounts will have had more time to tweet. The first round collected data for the initial core accounts (n = 52) and returned 231,918 unique accounts for further investigation. Collecting data for all 231,918 accounts would take around 240 days, which was considered impractical. Furthermore, the majority of these accounts will not be relevant to retrofit.

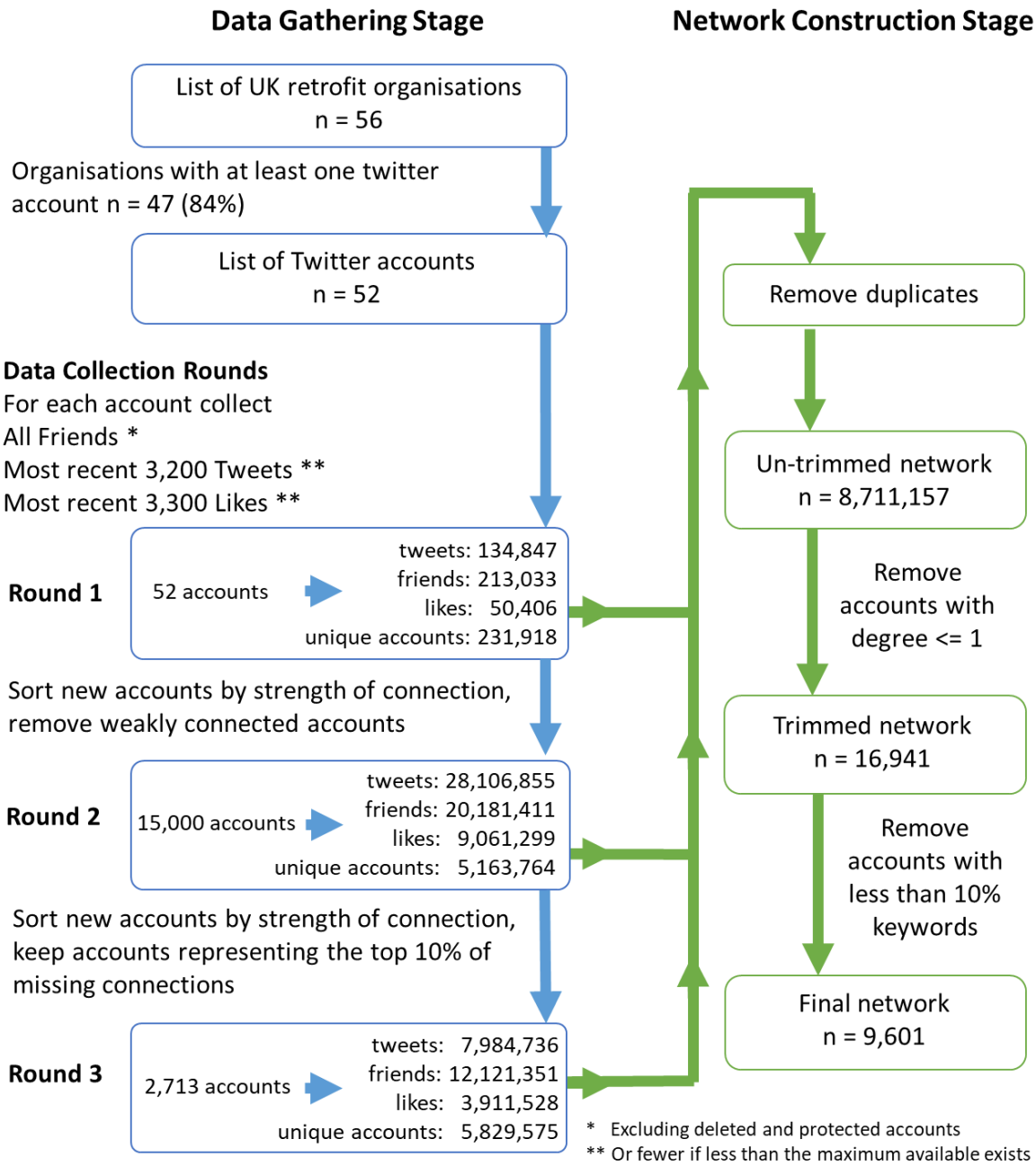

Figure 6: Outline of the data collection process

It was found that the majority of accounts (154,613 or 67%) had only one or two connections to the core accounts, as shown in Figure 7. The graph in Figure 7 shows that as the number of accounts is increased, the proportion of all connections rapidly becomes asymptotic with the one connection per account line. This is by definition the minimum rate of increase, as accounts with fewer than one connection cannot appear in the dataset. In the second round, data was then gathered for the first 15,000 accounts. 15,000 accounts were selected as a reasonable compromise, due to it ensuring that all accounts with more than three connections were included, while also limiting the total data collection time to under a month.

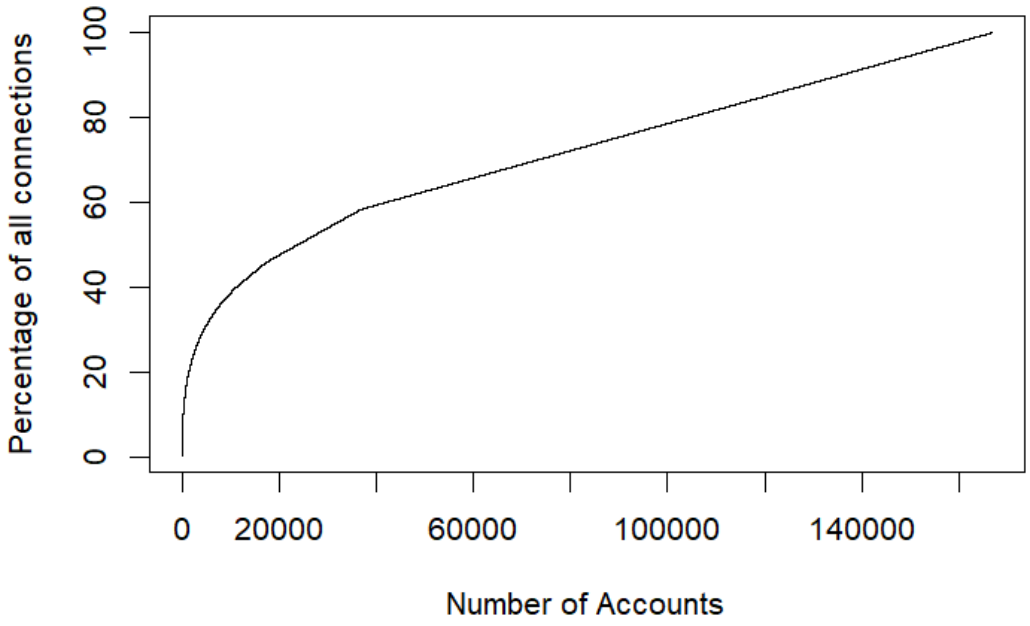

Figure 7: Cumulative percentage of all connections by the number of accounts. Notice the line become straight above 40,000 accounts representing only one additional connection per account.

The final round of data collection was a short round intended to identify any important accounts that we missed in the second round. As before, the accounts were sorted and examined. It was found that the missing accounts were a mix of off topic accounts (such as celebrities and charities not related to retrofit or the construction sector) and some housing sector individuals and business accounts. The vast majority of accounts had few connections to the rest of the network. Therefore, the missing accounts were sorted by the number of connections, and a short round of additional data collection was performed for the accounts representing the top 10% of missing connections. In total data for an additional 2,713 accounts was gathered.

3.3 Constructing a network

To convert the twitter data gathered into a network, the friends, likes, retweets, and mentions were summarised as connections from one account to another and then counted to give a single unweighted total number of interactions. Table 4 illustrates how the data collected from Twitter was summarised in preparation for constructing the network.

Table 4: Illustrative example of the interactions table

From	To	Friends	Likes	Retweets	Mentions	Total
jrf	fmbuilder	1	241	173	526	941
fmbuilder	jrf	1	16	21	27	65
jrf	UKGBC	0	1	3	1	5

The summary data, as shown in Table 4, was then passed to the igraph package (Csardi and Nepusz, 2016) to construct a network graph, which is discussed below. The network produced is very large (n = 8,711,157) but most of the twitter accounts on the network have very weak

connections to the rest of the network. It is therefore beneficial to trim the network down to a manageable and more relevant size.

3.4 Trimming the network

Many Twitter users mix professional and personal content in their Twitter usage. So this resulted in a mixture of retrofit specific and general interest accounts in the collected data. To ascertain the social structure of the UK retrofit sector it is necessary to remove their general interest accounts. Several common themes were observed in the types of non-retrofit account in the collected data:

- Celebrities and popular media (Films, TV shows, Football Teams);
- General interest news media and journalists (BBC, ITV, The Times)
- Politicians and Government agencies outside the area of retrofit (Theresa May, Ministry of Defence, Visit London)
- Prominent international individuals (Donald Trump, Mark Ruffalo)
- Charities and public campaigns (Gates Foundation, Oxfam, Help for Heroes)
- Business outside the construction sector (Virgin Trains, Costa Coffee, Amazon)

Without removing these types of accounts, false connections may be drawn about the retrofit industry. For example, a shared interest in Arsenal football club by two individuals in the construction sector is unlikely to indicate any retrofit-related communication between those two individuals. Conversely, individual politicians, journalists, or celebrities may perform an important role within the network as evangelists or communicators.

Since a large volume of text (tweets) has been gathered along with the network structure, text analysis of the tweets is possible. There is a wide range of methods for analysing text, but in this case, a simple keyword search is appropriate. The absence of keywords is a strong indicator that the account does not have a focus on retrofit and is therefore outside the remit of the research. As tweets are short, on average only 18 words long, it is difficult to assess the topic of a single tweet. A single tweet may contain very few words that can be used to identify its subject. For example:

> "From a distance Steel Farm looks like a traditional farmhouse. But inside it's something very different.\n\nRead more… https://t.co/RAws4j6Rzf" (Passive House Plus, 2017)

This tweet is highly relevant to retrofit but does not contain any specific retrofit words. In this case, the relevance of the tweet only becomes apparent when following the link to the associated article about retrofitting a farmhouse.

A typical account will have 2,600 tweets collected and thus has around 47,000 words of text to analyse. This provides a rich sample of text to look for keywords. Despite this large volume of text, a series of tweets is not the same as a single body of text of the same length. Twitter compels its users to be brief, and thus users are likely to omit and abbreviate words. In the extreme, an account may just have tweets such as “read this article” containing no useful keywords. Therefore,

the threshold for considering an account relevant based on its keywords should be low, and a wide range of possible keywords must be considered.

To identify the relevance of a twitter account to the retrofit sector the tweets of each account were searched for keywords. The keywords were produced first by a small group of experts in the field and then supplemented by common words and hashtags found within the Twitter data. Although this introduces a subjective element into the analysis it is necessary to specify a starting point, and impossible to scan though all twitter accounts to objectively find the optimal starting point prior to the data collection process. Table 6 in the appendices lists the keywords by 52 categories. The categories were used to distinguish between words that were exclusive to the domestic retrofit sectors, and those that overlap with other related fields such as construction and environmentalism.

Accounts were scored based on the occurrence of the keywords among their tweets for each of the 52 categories and in total. Accounts that had a keyword appear in at least 10% of their tweets were retained for further analysis. The 10% threshold excluded 7,340 accounts leaving 9,601 accounts within the network.

The final stage of trimming was to search for reciprocal contact between accounts. This method was based on (Huberman et al., 2008). As communication on Twitter is very easy, a single connection between accounts does not indicate significant communication. Therefore, connections were filtered to only allows those which are mutual (A connects to B, and B connects to A) and had occurred at least twice in each direction. This filtering technique reduced the number of connections between accounts from 2,897,813 to 1,034,136 a reduction of 64%. Removing these low importance connections made the network simpler and easier to analyse while preserving the most important connections.

4 Results

Cluster detection within the network was performed by the info map algorithm (Rosvall et al., 2009). This algorithm was chosen from the options available within the igraph package for its ability to account for both the direction and weight of connection within the network. Multiple distinct clusters were found with the overall modularity being 0.58. Figure 8 illustrates the structure of the network when each of the types of connection (friends, likes, retweets, and mentions) are combined and treated equally.

Within Figure 8, each of the 26 largest clusters has been assigned a unique colour. Accounts within small clusters (less than 50 members) are coloured black. For clarity, lines representing less than 50 connections been removed.

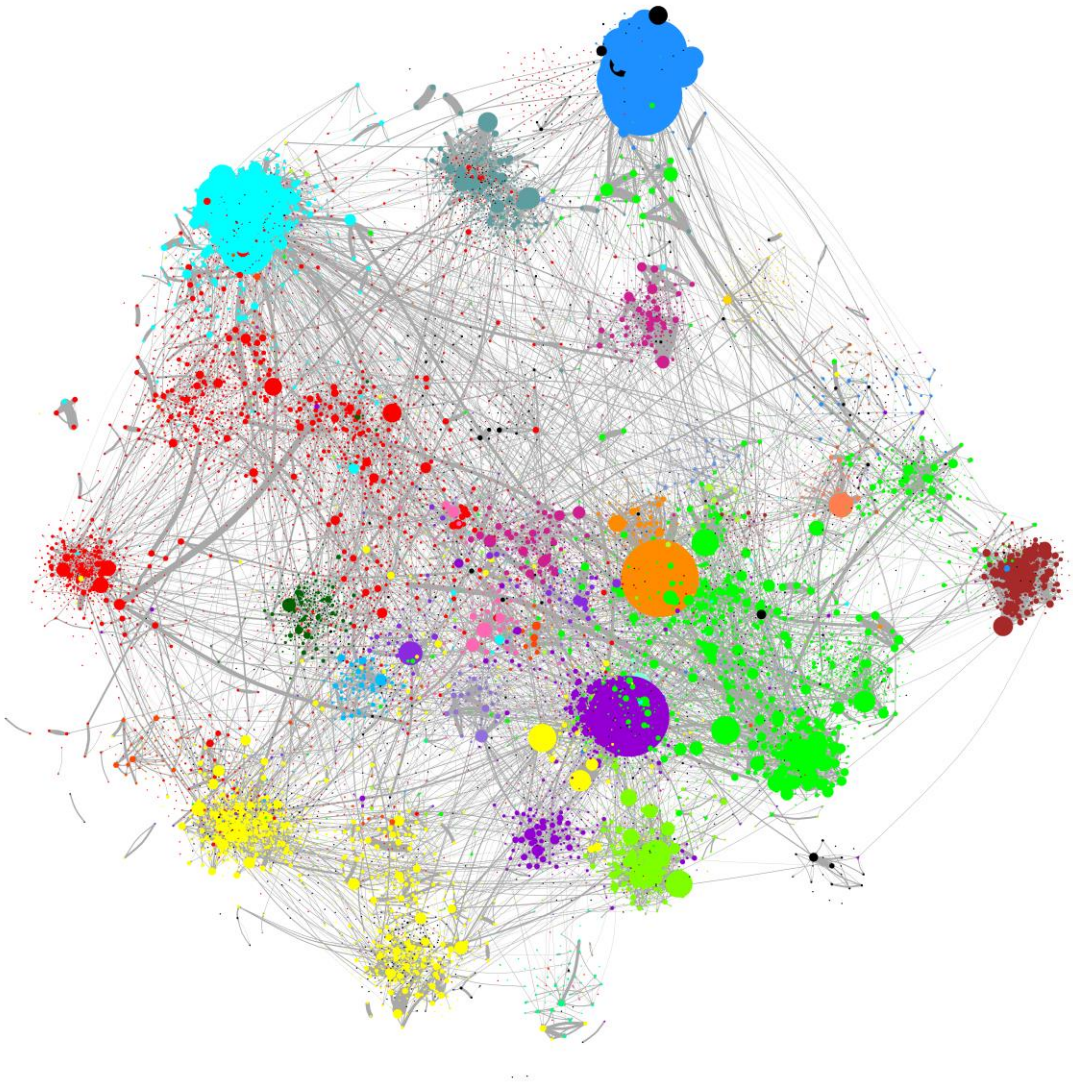

Figure 8: Plot of the whole twitter network, the layout is based on the DRL algorithm. The vertex size is based on the total strength (weighted number of connections) both in and out. Vertex colour shows the clusters as detected by the infomap algorithm with the largest 26 clusters given a unique colour. Vertices within smaller clusters are coloured black. Line width is proportional to the weight of the connection. For clarity line with a weight, less than 50 have been removed.

4.1 Clusters Identified

A summary of the largest clusters is shown in Figure 9 where each cluster identified in Figure 8 has been condensed to a single vertex. The cluster colouring and positioning is consistent with Figure 8. Within Figure 9, each cluster has been assigned a descriptive name to highlight the types of account within the cluster. The names have been chosen to reflect the keyword usage within the cluster and the prominent accounts within the cluster. However, the names are illustrative, not definitive. For example, the architecture cluster contains many architects and

architecture organisations, such as the Royal Institute of British Architects (RIBA). Within this cluster there are non-architects and there are architects within other clusters. Table 5 provides a short summary of the clusters including the top keywords used by members of the cluster.

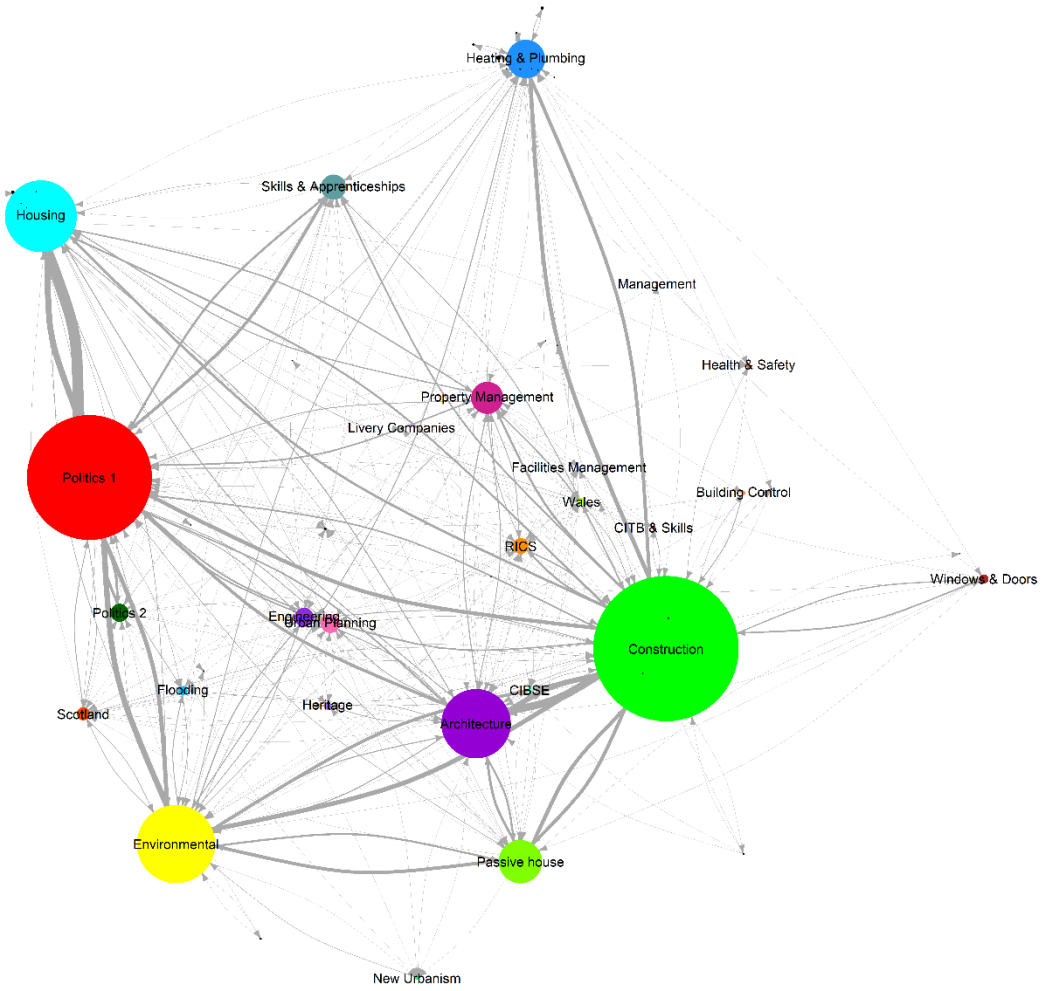

Figure 9: Simplified network showing the connections between the main clusters

name	***members***	***Top keywords***
Politics 1	*3507*	*brexit london labour eu women nhs m change home students*
Construction	*2272*	*construction bim building london design awards m home architecture*
Environmental	*1496*	*energy change carbon sustainability renewables sustainable eu climatechange renewable gas*
Housing	*1401*	*housing ukhousing homes home cihhousing london homelessness tenants residents house*
Architecture	*931*	*architecture design riba london architects house building housing awards architect*
Heating & Plumbing	*861*	*heating gas boiler pbplumber plumbing gasmangod vaillantuk energy home installermag*
Passive house	*672*	*passivhaus house energy building passive passivehouse design home homes housing*

Skills & Apprenticeships	*608*	*students apprenticeship apprenticeships skills courses apprentice careers fe job training*
Property Management	*510*	*london landlords home house housing landlord homes rent m tenants*
Windows & Doors	*343*	*windows doors door window home fitshow glazing solidorltd conservatory products*
Politics 2	*269*	*eu mp brexit conservatives labour mps parliament britain m fantastic*
RICS	*263*	*ricsnews rics construction london amandaclack building surveying women womenoffuture awards*
Engineering	*212*	*engineering ukmfg stem innovation women engineers students energy london congratulations*
Scotland	*191*	*scotland scottish glasgow edinburgh nicolasturgeon thesnp snp scotlands housing brexit*
Urban Planning	*152*	*housing london rtpiplanners homes cities thetcpa plans land brexit infrastructure*
Management	*139*	*pmot projectmanagement apm apmprojectmgmt change pmo projects agile stories training*
Heritage	*95*	*st building historicengland house london buildings listed conservation c spab*
Flooding	*94*	*flood flooding floodaware envagency info details floods customers thames m*
CITB & Skills	*80*	*construction citbuk citb citbscotland apprenticeship apprentice skills goconstructuk training apprenticeships*
Building Control	*78*	*labcuk building awards labc home labcawards homes labcwarranty construction cbuilde*
Wales	*78*	*wales yn construction welsh y cardiff ar o congratulations m*
CIBSE	*68*	*cibse energy building cibsewm engineering design london women buildings cibsejournal*
Facilities Management	*68*	*fm facman bifmawards workplace bifmuk thinkfm facilitiesmanagement facilities bifm ifma*
Health & Safety	*66*	*healthandsafety hse rospa ifsec home training childsafetyweek firefighters tips children*
New Urbanism	*61*	*cities placemaking ppsplacemaking trump design brenttoderian housing de building w*
Livery Companies	*55*	*london cityoflondon livery citylordmayor cityandlivery lord congratulations st marketors lordmayorsshow*

Table 5: Summary of the top clusters, show the top 10 keywords used by members of each cluster

The common theme within clusters varies significantly. Some clusters focus on a specific aspect of the retrofit sector (e.g. Heating & Plumbing, Windows & Door) while others are centred on a specific organisation (Charter Institute of Building Surveyors (CIBSE), Royal Institute of Charted Surveyors (RICS), and Construction Industry Training Board (CITB)). Several clusters focused on specific regions (Scotland and Wales). Finally, two clusters (Politics 1 & Politics 2) had a high prevalence of politicians, journalists, and government agencies with comparatively low usage of retrofit keywords.

It is not the purpose of this document to analyse these graphs in detail in relation to the social network of interest in retrofit, but some general comments can be made:

The method seems to work – it is possible to use computerised techniques to ‘reveal’ the social network for a topic of minority policy interest and communities and structure can be identified within the overall network.

5 Limitations of the method

It is important to acknowledge the limitations inherent in this method.

5.1 Twitter is not the offline world

As mentioned in section 2.4, Twitter is a biased subset of the population and there is a lack of evidence of how peoples' online behaviour differs from off-line behaviour. Therefore, any conclusions drawn from Twitter data cannot be generalised to the offline world without further evidence. However, the results of the Twitter analysis may still have value in understanding topics of minority interest if used correctly. The method may also provide a quick way to understand the structures of a social network as a precursor to off-line data collection.

5.2 Incomplete and uneven data collection

Due to limitations of the Twitter API, not all the relevant data could be collected for each account. This was a particular problem for older accounts or very active accounts where the total number of tweets, likes, and friends exceeded the maximum retrievable from the Twitter API. In these cases, the most recent data was collected. For some accounts, this means that as little as one percent of the tweets were gathered, while other accounts had all their tweets gathered. This imbalance is likely to reduce the reported importance of older and active Twitter accounts as they become relatively underrepresented in the network. Two possible methods to redress this imbalance would be to either limit the data to a specific period of time, e.g. only tweets in the last year, or to gather data repeatedly over a longer period.

Further data loss occurs when accounts are protected or deleted; these accounts are inaccessible from the Twitter API. However, they do appear in the network, a prominent example found in the gathered data was the Twitter accounts of the Department of Energy and Climate Change (DECC) and the Department of Business, Innovation and Skills (BIS) which were merged in July 2016 into the Department of Business, Energy, and Industrial Strategy (BIES).

Although it is possible through the Twitter API to gather followers of an account, this was not done in this method. Gathering followers would address the flaw show in Figure 5, where accounts in the network are undiscoverable due to the unidirectional nature of the search. Yet, introducing followers into the method would create two practical problems. Firstly, one of duplication of data collection, and secondly a massive increase in the volume of data collected per account. While many accounts have relatively few followers, some have millions of followers, requiring many hours to gather data for a single account.

Further research would be beneficial to ascertain if any of these limitations create systematic differences in the structure of the network.

5.3 Arbitrary choice of starting accounts

The method requires at least one starting account to begin the data collection. In this case, the choice of starting accounts was based on expert judgement thus is subject to bias. To mitigate the risk of bias a list of starting accounts was used. It was hoped that having multiple entry points into the network would reduce the risk of part of the network being missed. There is no way to verify if this is the case.

6 Conclusions and Further Work

This working paper has outlined a method to identify networks of interest in minority policy topics, in this case, the UK retrofit sector. The method has been shown to gather data as rapidly as possible allowing a reasonably large network ($n \approx 17{,}000$) to be collected in a practical period. Initial analysis of the resulting network suggests that meaningful sub-communities can be identified and that further analysis of Twitter data may yield useful insights into the retrofit sector. However, Twitter data must be used with care due to the lack of theoretical knowledge of how to interpret Twitter data and the lack of validation of how Twitter reflects the offline world.

The method also highlights the important of expert knowledge in defining a community to study, the as the selection of starting accounts may bias the results. This bias can be mitigated, although not entirely eliminated, by using a large list starting accounts which reflect the diversity of the community.

6.1 Further Work

The low cost of Twitter analysis, in comparison to other methods such as interviews and focus groups, makes it suitable for hypothesis testing as a precursor to conventional research techniques.

6.1.1 The packaging of the method

The functions listed in Appendix 1: Code Examples would benefit from being transferred to an R package that would make them easier to use, as packages are easier to install and contains analytical tools, rather than simple examples of how the analysis can be performed.

6.1.2 Dynamic detection of account relevance

Due to the rate limitation of the Twitter API, data collected on a non-relevant account has a high time cost; both in time spent not collecting other data and additional waiting time incurred. If a method could be developed to assess the relevance of the account based on an initial subset of the data, it may be possible to skip less relevant accounts during the data collection phase. Such a method could effectively increase the number of accounts collected without increasing the time taken for data collection.

6.1.3 Improved use of idle time

Due to the rate limiting of the Twitter API, the functions spend a significant proportion of the time idle. This represents a lost opportunity for the computer to perform other tasks. Some of the improvements suggested in this section could be incorporated into the data collection process to take advantage of any idle time.

6.1.4 A full analysis of the collected data

This working paper has only performed an initial analysis of the data collected. Further analysis is required before policy recommendations can be developed.

This page is intentionally left blank

References

Beguerisse-Díaz, M. 2013. Communities, roles, and informational organigrams in directed networks: the Twitter network of the UK riots. *arXiv preprint arXiv: …*., pp.1–29.

BEIS 2017. *The Clean Growth Strategy: Leading the way to a low carbon future* [Online]. London. Available from: https://www.gov.uk/government/publications/clean-growth-strategy.

Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M. and Wallach, D.S. 2015. Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Transactions on Internet Technology*. **15**(1), pp.1–24.

Bliss, C.A., Kloumann, I.M., Harris, K.D., Danforth, C.M. and Dodds, P.S. 2012. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*. **3**(5), pp.388–397.

Bright Planet 2018. Twitter Firehose vs. Twitter API: What's the difference and why should you care? [Accessed 18 August 2018]. Available from: https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/.

CCC 2018. Reducing UK emissions - 2018 Progress Report to Parliament - Committee on Climate Change. *Committee on Climate Change*. (June).

Cihon, P. and Yasseri, T. 2016. A Biased Review of Biases in Twitter Studies on Political Collective Action. , pp.1–10.

Conover, M., Ratkiewicz, J. and Francisco, M. 2011. Political polarization on twitter. *Icwsm*. **133**(26), pp.89–96.

Csardi, G. and Nepusz, T. 2016. The igraph software package for complex network research. *InterJournal*. **Complex Sy**, p.1695.

Gentry, J. 2016. twitteR. , pp.1–32.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y. 2014. Assessing the bias in samples of large online networks. *Social Networks*. **38**(January), pp.16–27.

Grabowicz, P.A., Ramasco, J.J., Moro, E., Pujol, J.M. and Eguiluz, V.M. 2012. Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media Y. Moreno, ed. *PLoS ONE*. **7**(1), p.e29358.

Granovetter, M.S., American, T. and May, N. 1973. The Strength of Weak Ties The Strength of Weak Ties1. *The American Journal of Sociology*. **78**(6), pp.1360–1380.

HMSO 2008. Climate Change Act 2008. , p.101.

Huberman, B.A., Romero, D.M. and Wu, F. 2008. Social networks that matter: Twitter under the microscope.

International Energy Agency 2017. *World Energy Outlook 2017* [Online]. Paris. Available from: http://www.iea.org/media/weowebsite/2017/Chap1_WEO2017.pdf.

IPCC 2014. *Climate change 2014: Mitigation of climate change*. Geneva.

Janda, K.B. and Parag, Y. 2013. A middle-out approach for improving energy performance in buildings. *Building Research and Information*. **41**(1), pp.39–50.

Killip, G. 2013. Products, practices and processes: Exploring the innovation potential for low-carbon housing refurbishment among small and medium-sized enterprises (SMEs) in the UK construction industry. *Energy Policy*. **62**, pp.522–530.

Killip, G., Fawcett, T. and Janda, K.B. 2014. Innovation in low-energy residential renovation: UK and France. *Proceedings of the Institution of Civil Engineers - Energy*. **167**(3), pp.117–124.

Kwak, H., Lee, C., Park, H. and Moon, S. 2010. What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors. , pp.591–600.

Mlecnik, E., Straub, A. and Haavik, T. 2018. Collaborative business model development for home energy renovations. *Energy Efficiency*. (Eu 2011), pp.1–16.

Myers, S.A., Sharma, A., Gupta, P. and Lin, J. 2014. Information network or social network? *In*: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* [Online]. New York, New York, USA: ACM Press, pp.493–498. Available from: http://dl.acm.org/citation.cfm?doid=2567948.2576939.

Owen, A., Mitchell, G. and Gouldson, A. 2014. Unseen influence-The role of low carbon retrofit advisers and installers in the adoption and use of domestic energy technology. *Energy Policy*. **73**, pp.169–179.

Passive House Plus 2017. Twitter. *Twitter*. [Online], p.1. Available from: https://twitter.com/i/web/status/880751623563350016.

Patrick, J., Killip, G., Brand, C., Augustine, A. and Eyre, N. 2014. *Oxfordshire's Low Carbon Economy* [Online]. Oxford. Available from: http://www.eci.ox.ac.uk/research/energy/downloads/olce-report-oct2014.pdf.

Rayner, S. 2010. How to eat an elephant: A bottom-up approach to climate policy. *Climate Policy*. **10**(6), pp.615–621.

Rosvall, M., Axelsson, D. and Bergstrom, C.T. 2009. The map equation. *The European Physical Journal Special Topics*. **178**(1), pp.13–23.

Skea, J., Ekins, P., Winskel, M., Howard, D., Eyre, N. and Hawkes, A. 2009. *Making the transition to a secure and low-carbon energy system: synthesis report*.

Sloan, L., Morgan, J., Burnap, P. and Williams, M. 2015. Who Tweets ? Deriving the Demographic Characteristics of Age , Occupation and Social Class from Twitter User Meta-Data. , pp.1–20.

Stafford, A., Gorse, C. and Shao, L. 2011. The Retrofit Challenge: Delivering Low Carbon Buildings. *Research Insights into Building Retrofit for the UK*., pp.1–32.

Twitter 2018. Docs. [Accessed 3 April 2018]. Available from: https://developer.twitter.com/en/docs.

Valenzuela, S., Correa, T. and Zúñiga, H.G. De 2018. Ties , Likes , and Tweets : Using Strong and Weak Ties to Explain Differences in Protest Participation Across Facebook and Twitter Use Ties , Likes , and Tweets : Using Strong and Weak Ties to Explain Differences in Protest Participation Across Facebook a. *Political Communication*. **35**(1), pp.117–134.

WBCSD 2009. *Transforming the market: energy efficiency in buildings*. Geneva.

Weng, L. 2014. Information Diffusion on Online Social Networks. . (April), pp.1–168.

Yang, Z., Algesheimer, R. and Tessone, C.J. 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*. **6**(August).

7 Appendices

7.1 Appendix 1: Code Examples

7.1.1 Rate Limiting

```
rate.wait <- function(user, type = c("Friends","Followers","Tweets","Favorites")){
  type.fixed <- paste0(type, paste(rep(" ",9 - nchar(type)), collapse = ''))
  if(type == "Friends"){
    nreq <- ceiling(user$friendsCount / 5000)
    resource <- "/friends/ids"
  }else if(type == "Followers"){
    nreq <- ceiling(user$followersCount / 5000)
    resource <- "/followers/ids"
  }else if(type == "Tweets"){
    nreq <- 1
    resource <- "/statuses/user_timeline"
  }else if(type == "Favorites"){
    nreq <- ceiling(user$favoritesCount / 200)
    resource <- "/favorites/list"
  }else{
    message(paste0("Unknown Type of wait request, ",type,", in function rate.wait" ))
    stop()
  }
  limit <- getCurRateLimitInfo()
  limit <- limit[limit$resource == resource,]
  lim <- as.numeric(limit$limit[1])
  rem <- as.numeric(limit$remaining[1])
  if(nreq <= rem){
  }else if((nreq > rem) & (nreq <= lim)){
    wait <- limit$reset[1] - Sys.time()
    wait <- as.integer(as.numeric(wait, units = "secs")) + 5
    if(wait < 0){wait <- 1}
    message(paste0(Sys.time()," ",type.fixed,": Waiting for ",round(wait/60,1)," minutes"))
    Sys.sleep(wait)
  }else if(nreq > lim){
    message(paste0(Sys.time()," ",type,": Request of ",nreq," exceeds the maximum of ",lim,". Making an attemp"))
    message("Waiting will increase the number of results returned")
    wait <- limit$reset[1] - Sys.time()
    wait <- as.integer(as.numeric(wait, units = "secs")) + 5
    if(wait < 0){wait <- 1}
    message(paste0(Sys.time()," ",type.fixed,": Waiting for ",round(wait/60,1)," minutes"))
    Sys.sleep(wait)
  }else{
    warning(paste0(Sys.time()," ",type,":Unknown number of requests has occured: nreq: ",nreq," reamining: ",rem," limit: ",lim))
    stop()
  }
}
```

7.1.2 Data Collection

```
get.data <- function(x, type = c("Friends","Followers","Tweets","Favorites")){
  start.time <- Sys.time()
  type.fixed <- paste0(type, paste(rep(" ",9 - nchar(type)), collapse = ''))
  if(class(x)[1] == "user"){
    user <- x
    id <- x$screenName
  }else if(class(x)[1] == "character"){
    id <- x
    user <- getUser(id)
  }else{
    message(paste0(Sys.time()," ",type.fixed,": Unknown type of input ",class(x)[1], " in function get.friends"))
    stop()
  }
  if(type == "Friends"){
    proceed <- (user$friendsCount == 0)
  }else if(type == "Followers"){
    proceed <- (user$followersCount == 0)
  }else if(type == "Tweets"){
    proceed <- (user$statusesCount == 0)
  }else if(type == "Favorites"){
    proceed <- (user$favoritesCount == 0)
  }else{
    message(paste0(Sys.time()," Unknown type: ",type,", in function get.data"))
    stop()
  }
  if(user$protected | proceed){
    message(paste0(Sys.time()," ",type.fixed,": User ",id," has a protected account or no ",type ,"s to collect so skipping"))
  }else{
    if(type != "Favorites"){
      rate.wait(user = user, type = type)
    }
    if(type == "Friends"){
      friends <- try(user$getFriends())
      if(class(friends) == "try-error"){
        friends <- NULL
        data.df <- NULL
        message(paste0(Sys.time()," Friends: Unable to find account ",user$screenName," moving to next"))
      }else{
        friends <- friends[sapply(friends,class)=="user"]
        data.df <- do.call("rbind", lapply(friends, as.data.frame))
        data.df$friendof <- id
        data.df$followerof <- NA
      }
      data.total <- user$friendsCount
    }else if(type == "Followers"){
      follower <- try(user$getFollowers())
      if(class(follower) == "try-error"){
        follower <- NULL
        data.df <- NULL
        message(paste0(Sys.time()," Followers: Unable to find account ",user$screenName," moving to next"))
      }else{
        follower <- follower[sapply(follower,class)=="user"]
        data.df <- do.call("rbind", lapply(follower, as.data.frame))
        data.df$friendof <- NA
        data.df$followerof <- id
      }
      data.total <- user$followersCount
    }else if(type == "Tweets"){
      tweets <- try(userTimeline(id, n = 3200, includeRts=TRUE, excludeReplies=FALSE))
      if(class(tweets) == "try-error"){
        tweets <- NULL
        data.df <- NULL
        message(paste0(Sys.time()," Tweets: Unable to find account ",user$screenName," moving to next"))
      }else{
        data.df <- twListToDF(tweets)
        data.df <- data.df[!duplicated(data.df$id),]
      }
      data.total <- user$statusesCount
```

```
    }else if(type == "Favorites"){
      data.df <- get.data.favorites(user)
      data.total <- user$favoritesCount
    }else{
      message(paste0(Sys.time()," Unknown type: ",type,", in function get.data"))
      stop()
    }
    end.time <- Sys.time()
    message(paste0(Sys.time()," ",type.fixed,": ",round(nrow(data.df)/data.total*100,1),"% at
",round(nrow(data.df)/as.numeric(difftime(end.time,start.time,units = "secs")),0) ,"/sec for ",id))
    return(data.df)
  }
}
```

7.1.3 Data Collection (Likes / Favourites)

```
get.data.favorites <- function(user){
  start.time <- Sys.time()
  nreq <- ceiling(user$favoritesCount / 200)
  favs.list <- list()
  nreq.todo <- min(c(nreq,33))
  for(a in seq(from = 1, to = nreq.todo)){
    limit <- getCurRateLimitInfo()
    limit <- limit[limit$resource == "/favorites/list",]
    rem <- as.numeric(limit$remaining[1])
    if(rem <= 0){
      wait <- as.integer(as.numeric(limit$reset[1] - Sys.time(), units = "secs")) + 1
      if(wait < 0){wait <- 1}
      message(paste0(Sys.time()," Favorites: Waiting for ",round(wait/60,1)," minutes due to : Used up allowance in a loop"))
      Sys.sleep(wait)
      rm(limit,rem,wait)
    }else{
      rm(limit,rem)
    }
    if(a == 1){
      minid <- NULL
    }
    favs <- try(user$getFavorites(n = 100, max_id = minid), silent = T)
    if(class(favs) == "try-error"){
      favs <- NULL
      message(paste0(Sys.time()," Favorites: Unable to find account ",user$screenName," during loop number ",a," moving to next
loop"))
    }else{
      if(length(favs) > 1 | (a == 1 & length(favs) == 1) ){
        favs <- twListToDF(favs)
        favs.list[[a]] <- favs
        minid <- min(as.double(favs$id)) - 1
      }else{
        break
      }
    }
    rm(favs)
  }
  favs.list <- favs.list[lapply(favs.list,length)>0]
  favs.df <- do.call("rbind",favs.list)
  if(!is.null(favs.df)){
    favs.df <- favs.df[!duplicated(favs.df$id),]
    favs.df$favOf <- user$screenName
  }
  end.time <- Sys.time()
  return(favs.df)
}
```

7.1.4 Data Collection (Users)

```
get.users <- function(ids, output = c("data.frame","list")){
  start.time <- Sys.time()
  accounts.list <- list()
  ids <- unique(ids)
  nreq <- length(ids)
  limit <- getCurRateLimitInfo()
  limit <- limit[limit$resource == "/users/show/:id",]
  lim <- as.numeric(limit$limit[1])
  rem <- as.numeric(limit$remaining[1])
  if(rem == 0){
    wait <- as.integer(as.numeric(limit$reset[1] - Sys.time(), units = "secs")) + 5
    if(wait < 0){wait <- 1}
    message(paste0(Sys.time()," Users: Waiting for ",round(wait/60,1)," minutes due to starting with no allowance"))
    Sys.sleep(wait)

    limit <- getCurRateLimitInfo()
    limit <- limit[limit$resource == "/users/show/:id",]
    lim <- as.numeric(limit$limit[1])
    rem <- as.numeric(limit$remaining[1])
  }
  loops <- list()
  if(nreq >= rem){
    loops[[1]] <- rem
  }else{
    loops[[1]] <- nreq
  }
  loops.extra <- ceiling((nreq - rem)/lim)
  if(loops.extra >= 1){
    left <- nreq - rem
    for(a in 1:loops.extra){
      if(left >= lim){
        loops[[a + 1]] <- lim
        left <- left - lim
      }else{
        loops[[a + 1]] <- left
      }
    }
  }
  loops <- unlist(loops)
  for(b in 1:length(loops)){
    for(c in 1:loops[b]){
      if(b == 1){
        idno <- c
      }else{
        idno <- c + sum(loops[1:(b-1)])
      }
      user <- try(getUser(ids[idno]), silent = T)
      if(class(user) == "try-error"){
        user <- NULL
        message(paste0(Sys.time()," Users: Unable to find account ",ids[idno]," moving to next account"))
      }else{
        if(output == "data.frame"){
          user <- user$toDataFrame()
          accounts.list[[idno]] <- user
        }else if(output == "list"){
          accounts.list[[idno]] <- user
        }else{
          message("Unknown Output Type")
          stop()
        }
      }
    }
    if(b != length(loops)){
      limit <- getCurRateLimitInfo()
      limit <- limit[limit$resource == "/users/show/:id",]
      wait <- as.integer(as.numeric(limit$reset[1] - Sys.time(), units = "secs")) + 5
      if(wait < 0){wait <- 1}
      message(paste0(Sys.time()," Users: Waiting for ",round(wait/60,1)," minutes"))
```

```
      Sys.sleep(wait)
    }
  }
  if(output == "data.frame"){
    accounts.df <- do.call("rbind",accounts.list)
    accounts.df <- accounts.df[!duplicated(accounts.df$id),]
    end.time <- Sys.time()
    message(paste0(Sys.time(),                "              ",nrow(accounts.df),"/",length(ids),              "              users                    @
",round(nrow(accounts.df)/as.numeric(difftime(end.time,start.time,units = "secs")),1) ," users/second"))
    return(accounts.df)
  }else if(output == "list"){
    accounts.list <- accounts.list[lapply(accounts.list,length)>0]
    end.time <- Sys.time()
    message(paste0(Sys.time(),                "              ",length(accounts.list),"/",length(ids),              "              users                    @
",round(length(accounts.list)/as.numeric(difftime(end.time,start.time,units = "secs")),1) ," users/second"))
    return(accounts.list)
  }
}
```

7.1.5 Data Collection (Wrapper Function)

```
get.SNAdata <- function(ids, temp.fld, batch.start = 1, trim = FALSE){
  start.time <- Sys.time()
  ids <- unique(ids)
  nreq <- length(ids)
  loops <- list()
  lim = 50
  if(nreq >= lim){
    loops[[1]] <- lim
  }else{
    loops[[1]] <- nreq
  }
  loops.extra <- ceiling((nreq - lim)/lim)
  if(loops.extra >= 1){
    left <- nreq - lim
    for(a in 1:loops.extra){
      if(left >= lim){
        loops[[a + 1]] <- lim
        left <- left - lim
      }else{
        loops[[a + 1]] <- left
      }
    }
  }
  loops <- unlist(loops)
  message(paste0(Sys.time()," To get ",nreq," accounts will require ",length(loops)," batches of up to ",lim))
  accounts.list <- list()
  friends.list <- list()
  favorites.list <- list()
  tweets.list <- list()
  for(b in seq(from = batch.start, to = length(loops))){
    if(b == 1){
      idnos <- 1:loops[1]
    }else{
      idnos <- (sum(loops[1:(b-1)]) + 1) : (sum(loops[1:(b-1)]) + loops[b])
    }
    message(paste0(Sys.time()," Doing batch ",b,": Getting account details"))
    batch.accounts <- get.users(ids[idnos], output = "list")
    batch.accounts.list <- list()
    for(e in seq(1,length(batch.accounts))){
      sub <- batch.accounts[[e]]
      sub <- sub$toDataFrame()
      batch.accounts.list[[e]] <- sub
    }
    batch.accounts.df <- do.call("rbind",batch.accounts.list)
    accounts.list[[b]] <- batch.accounts.df
    tasks <- list(
      job1 = function() lapply(batch.accounts, get.data, type = "Friends"),
      job2 = function() lapply(batch.accounts, get.data, type = "Favorites"),
      job3 = function() lapply(batch.accounts, get.data, type = "Tweets")
    )
    if(!dir.exists(paste0("twitterlog"))){
      dir.create(paste0("twitterlog"))
    }
    message(paste0(Sys.time()," Doing batch ",b,": Starting Cluster"))
    cl <- makeCluster( length(tasks), outfile = paste0("twitterlog/parlog-",b,"-",Sys.Date(),".txt") )
    clusterExport(cl=cl, varlist=c("batch.accounts"), envir=environment())
    clusterEvalQ(cl, {library(twitteR); source("functions.R"); source("secrets.R")})
    out <- clusterApply(cl,tasks,function(f) f())
    stopCluster(cl)
    message(paste0(Sys.time()," Doing batch ",b,": Cleaning Results"))
    out1 <- out[[1]]
    out2 <- out[[2]]
    out3 <- out[[3]]
    out1 <- out1[lapply(out1,length)>0]
    out2 <- out2[lapply(out2,length)>0]
    out3 <- out3[lapply(out3,length)>0]
    out1.bind <-  try(do.call("rbind",out1))
    if(class(out1.bind) != "data.frame"){
```

```
      message(paste0(Sys.time()," Out1 Rbind Fail saving result"))
      saveRDS(out1,paste0(temp.fld,"/RbindFail-out1-",b,"-",Sys.Date(),".Rds"))
      out1.bind <- NULL
    }else{
      if(trim){
        out1.bind <- out1.bind[,c("description","statusesCount","followersCount","favoritesCount","friendsCount","name",
                                  "created","protected","verified","screenName","location","lang","id",
                                  "listedCount","friendof","followerof")]
      }
    }
    out2.bind <-  try(do.call("rbind",out2))
    if(class(out2.bind) != "data.frame"){
      message(paste0(Sys.time()," Out2 Rbind Fail saving result"))
      saveRDS(out2,paste0(temp.fld,"/RbindFail-out2-",b,"-",Sys.Date(),".Rds"))
      out2.bind <- NULL
    }else{
      if(trim){
        out2.bind <- out2.bind[,c("text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID",
                                  "statusSource","screenName","retweetCount","isRetweet","retweeted","favOf")]
      }
    }
    out3.bind <-  try(do.call("rbind",out3))
    if(class(out3.bind) != "data.frame"){
      message(paste0(Sys.time()," Out3 Rbind Fail saving result"))
      saveRDS(out3,paste0(temp.fld,"/RbindFail-out3-",b,"-",Sys.Date(),".Rds"))
      out3.bind <- NULL
    }else{
      if(trim){
        out3.bind <- out3.bind[,c("text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID",
                                  "statusSource","screenName","retweetCount","isRetweet","retweeted")]
      }
    }
    friends.list[[b]] <- out1.bind
    favorites.list[[b]] <- out2.bind
    tweets.list[[b]] <- out3.bind
    rm(out,out1,out2,out3,out1.bind,out3.bind,out2.bind)
    if(!is.null(temp.fld)){
      saveRDS(friends.list,paste0(temp.fld,"/FriendsList-",Sys.Date(),"-bs-",batch.start,".Rds"))
      saveRDS(favorites.list,paste0(temp.fld,"/FavoritesList-",Sys.Date(),"-bs-",batch.start,".Rds"))
      saveRDS(tweets.list,paste0(temp.fld,"/TweetsList-",Sys.Date(),"-bs-",batch.start,".Rds"))
      saveRDS(accounts.list,paste0(temp.fld,"/AccountsList-",Sys.Date(),"-bs-",batch.start,".Rds"))
    }
  }
  end.time <- Sys.time()
  message(paste0(Sys.time(), " data gathered for ",length(ids), " users "))
  return(NULL)
}
```

7.2 Appendix 2: Keyword Groups

Table 6: List of keywords used to identify retrofit relevant twitter accounts

Group	Keywords
affordable	affordable affordablehousing affordability
architect	architecture architects architect architectsjrnal architectural architecturaljobs architectmark architectmag
bed	bed bedroom beds
bim	bim bimireland bimm bimsme bimstore bimcrunch bimacademy bimopenmic bimshowlive bimsummit bimgcs bimplus bimprospects bimsmeawards
build	building build buildings built builders builder
carbon and climate	carbon climatechange emissions climateaction resilience carbonbrief climategroup climateweek changeclimatechange climatehome citiesresearch resilient citiesclimate carbonbubble climatecentral carbonfix climateclg climatereality emission
CIBSE	cibse cibsewm cibsejournal cibseawards
CIH	cihhousing cih cihfutures cihcymru cihevents cihnw cihscotland cihtbc cihse cihsw cihpolicy cihne cihwestmidlands cihscot cihneconf ciheast
construction	construction standards contractors contractor
design	design designed designer designers designing
doors and windows	windows doors door window glazing glazingblogger doubleglazing glazed windownewsuk glasstimes windowsactive glassnewsmag windowwidgets glazeriteltd
eco	eco ecobuildnow ecobuild ecologicalbuild ecotec ecodistricts ecofit ecofys
economy	economy economic economies economics
efficiency	efficiency energyefficiency efficient energyefficient
electric	electricity electric electrical
energy	energy energystorage energyunion energyutilities energyeurope energyhour energyukcomms energydesk energylivenews energyinstitute
engineering	engineering engineers engineer
environment	environment environmental
flood	flood flooding floods floodaware floodmary floodandcoast floodre
floor	floor flooring
fuel	fuel fuels
gas	gas boiler gasmangod irbheating boilers gassaferegister gassafetyweek gassafepete gaskellmike carbonmonoxide gassafe gassafeglasgow gaschatgroup gaschattour gasboilersalton gassafety gasappuk
green	greenbuild greenbuilding greeninfrastructure greenhouse greendeal
heating	heating heat heatingconsult heatingcontrols heatingyourhome heater heaters
homeless	homelessness homeless homelessnessreductionbill
house	housing home house homes ukhousing housingday homesforbritain houses ukhousingfast homesforwales homeowners housingitguy housingcrisis housemarkltd housingjobs housingawards households housebuilding housetorent housingsubutcher housingfirst resourcehousing houseplanhelp ukhousingawards housingbill homeimprovement homesproperty housingwhitepaper household housingmagazine houseexchange housingex
infrastructure	infrastructure insidehousing
installer	installation installer installermag installed installers installersfirst installing installations installershow installersunion
insulation	insulation insulated
JRF	jrfuk jrfbrian jrf jrfuks
land	land landaid
landlord	landlords landlord landlordtweets landlordref landlordinshow nationalandlord landlordaction landlordzone landlordshow
natfed	natfednews natfeddavid natfedevents natfed natfenawards natfedkatie natfedclaire
neighbourhood	neighbourhood neighbourhoods neighbourhoodplanning
passivehouse	passivhaus passive passivehouse passivhaustrust passivehousebb passivhausnews passivelogical passiveacademy passivehousecal phplusmag phpp
plumbing	pbplumber plumbing bathroom bath plumbingac bathrooms pbmag plumbpal plumbpalproduct pbmmagazine phpi
poverty	poverty ukpoverty fuelpoverty
property	propertynews propertyweek propertyshe propertydanh propertyhour ukproperty propertywire

renewables	renewables renewable pv renewableenergy solarpv solarpowerport renewableuk solarenergy solarpowereu solarcentury solaraid solarimpulse solareditor renews
rent	rent tenants tenant rental rents rented renting tenancy renters tenancydeposits tenancies tenure
retrofit	renovation retrofit repairs repair replacement restoration replacing replaced retrofitawards
residential	residents residential residence
RIBA	riba ribaj ribaawards ribanorthwest ribalondon ribanortheast ribaarchitect ribasoutheast ribaeastmidland ribawestmids ribayorkshire ribanorth ribabookshops ribas ribacomps
RICS	ricsnews rics ricsawards ricseurope ricsmatrics ricschiefexec ricsapc ricsamericas ricssota ricswales ricsnorth ricsrecruit ricsscotland ricsfutures ricseastmids ricsresi ricssouth ricscpd ricsmodus
roof	roof roofing roofs rooflights roofingtoday roofers greenroofs roofingawards rooftophousing greenroof ukroofingawards rooflight roofer rooftop
skills	training skills apprenticeships apprenticeship apprentice apprentices citbuk citb citbscotland citbwales traintradeskills
smart	smart smartcities
surveying	surveying surveyors surveyor surveyingthefuture surveys
sustainability	sustainable sustainability
tax and benefits	benefits righttobuy bedroomtax
timber	timber wood timbercomposite wooden timberexpo woodawards
trades	tradestalk trades tradetalk tradesmen