Lost Umbrellas:
Bias and the Right to be Forgotten in Social Media Research

Rebekah Tromble

Daniela Stockmann

Forthcoming in Michael Zimmer and Katharina Kinder- Kurlanda, eds, In Press.

Internet Research Ethics for the Social Age: New Cases and Challenges.

New York: Peter Lang Publishers.

Since the European Court of Justice handed down its ruling in the 2014 *Costeja* case—finding that Google and other search engine operators must consider requests made by individuals to remove links to websites that contain the requesting party's personal information—scholars, policymakers, legal practitioners, media commentators, and corporate representatives around the globe have been vigorously debating the so-called "right to be forgotten." In the American context, many worry that recognizing such a right would undermine the First Amendment's protections for freedom of speech and press. In the European Union, a renamed "right to erasure" is expected to become law as part the EU's General Data Protection Regulation in 2016. The right to erasure "prevent[s] the indefinite storage and trade in electronic data, placing limits on the duration and purpose for which businesses" can retain such data (Tsesis, 2014, 433) and holds that individuals may request the deletion of data when those data have become irrelevant, are inaccurate, or cause the individual harm that is not outweighed by a public benefit in retaining the data (Koops, 2011).

Though most of the discussion surrounding the right to be forgotten and right to erasure has focused on the limits and responsibilities of corporate and media data "controllers," internet users' basic right to remove and have removed content they personally generate—including content that they believe may have a detrimental effect on how they are publicly viewed—also needs to be taken seriously by scholars conducting internet research. At a minimum, the right to be forgotten points to important ethical concerns about research subjects' privacy, as well as how and when a subject's consent is given and withdrawn. Indeed, if we accept the common argument that formal consent need not be obtained from research subjects whom have made their content entirely open to the public, the corollary would suggest that we have a responsibility to delete their data from our datasets when it is has been removed from the public domain.

And yet to do so could undermine the validity and reliability of social scientific research findings, introducing bias and undercutting reproduction and replication efforts. Indeed, respecting and observing the right to be forgotten has the potential to hamper ongoing movements for greater social science data sharing and transparency. Hoping to increase the accessibility of publicly funded research, thwart data falsification, and improve the reproducibility and replicability of social science studies, researchers and policymakers have vowed to make data even more widely available. Thus, we face a dilemma: Do we protect the rights of research subjects by deleting their data when it is no longer in the public domain? Or do we safeguard the scientific process and the integrity of our research results—sharing data widely and making the right to erasure effectively impracticable?

In order to understand and address this dilemma, we first need a better grasp of just how serious the implications of honoring the right to erasure would be for social science research. That is, we need a clearer understanding of whether and to what extent inferences might be biased, and basic scientific replicability undermined, if deleted internet content were indeed removed from our datasets. To this end, we examine two Twitter datasets related to the 2014 Hong Kong protests, often referred to as the "umbrella revolution" or "umbrella movement." We collected these data from Twitter's historical archive using the same search parameters at two points in time—in December 2014, just as the Hong Kong protests were winding down, and one year later in December 2015—and we use these datasets to assess the number of tweets deleted, as well as how these deletions impact social network metrics derived from the data.

As a case study, the umbrella movement presents an excellent opportunity to gauge, in concrete and practical ways, how the right to erasure might impact a large, growing, and influential body of work on the use of social media by social movement activists (cf. Gonzalez-

Bailon et al, 2011; Hanna, 2013; Harrigan et al, 2012; Tremayne, 2014). The Hong Kong protests represent a case in which the subjects being studied are likely to have compelling reasons to exercise their right to be forgotten. Though somewhat freer to express their views than are those in mainland China, Hong Kong residents have reason to be concerned about state censorship and repression and may wish to delete content to avoid monitoring, detention, or other forms of state control. The Hong Kong protests therefore represent exactly the type of case that should stimulate ethical concerns among internet researchers.

In laying out this analysis, we begin by offering a more detailed discussion of the umbrella movement, elucidating its context and developments. We then present a brief overview of Twitter, its archive, and the rules the company lays out for data use, including data deletion. Next, we provide a short description of the methods used to collect our Twitter data before moving on to an analysis of the differences between our two datasets and a discussion of the implications of these differences for social scientific research.

The Hong Kong Umbrella Movement

Protest erupted in Hong Kong on September 22, 2014 in reaction to a decision by the National People's Congress (NPC) of the People's Republic of China regarding electoral reform for the Chief Executive in Hong Kong. Currently, the Chief Executive is chosen by an election committee.[1] For the 2017 elections, the NPC decided that voters should be able to choose from a

[1] For details of the composition of the election committee, see Annex I of Basic Law at http://www.basiclaw.gov.hk/en/basiclawtext/images/basiclaw_full_text_en.pdf, accessed on April 16, 2016.

list of two or three candidates selected by the election committee and that each nominee would be eligible to run if he or she secured the support of more than 50% of that committee.[2] Critics argued that the election committee overrepresented the interests of Beijing and that without democratizing the selection of the election committee itself, the popular vote for the Chief Executive constituted mere window dressing. Because the Basic Law of the Hong Kong Special Administrative Region expressed the ultimate aim of selecting the Chief Executive by universal suffrage (upon nomination by a broadly representative election committee in accordance with democratic procedures), protesters called for Beijing to fulfil its promise to implement genuine universal suffrage in this process. Supporters of the decision argued that the letter of the law leaves room for interpretation and does not specify the timing of gradual electoral reforms.

During the protests, students and other citizens occupied a central square in Hong Kong, often referred to simply as "Central," as well as a few shopping streets. The occupation and protests came to be known as the "umbrella movement" or "umbrella revolution" after the umbrellas pro-democracy protesters held up as a protection against tear gas fired by police. Yellow ribbons also emerged as a symbol for peace worn by supporters and were seen fluttering in the city to condemn the use of tear gas and violence by the Hong Kong police.

But not everyone in Hong Kong agreed with the umbrella movement and some began displaying blue ribbons to support the authorities and the police (the latter of whom wear blue uniforms). Blue-ribbon supporters accused student protesters of engaging in violent protests and

[2] http://news.xinhuanet.com/politics/2014-08/31/c_1112298240.htm, accessed on April 16, 2016; http://www.bbc.com/news/world-asia-china-27921954, accessed on April 16, 2016.

of severely disrupting social order.[3] The blue ribbon counter-movement also took to the streets, and numerous clashes took place between yellow and blue ribbon supporters until the occupation ended on December 15, 2014. Ultimately, the umbrella movement protests failed to secure revisions to the NPC Standing Committee's electoral procedures.

While the Chinese government avoided direct contact with the protesters, it kept a close eye on how Hong Kong officials handled the protests and sought to direct the response from behind closed doors.[4] As early as September 28, the Propaganda Department, State Council Information Office, and related institutions issued directives to strictly manage interactive media and delete all harmful information regarding "occupy central."[5] Words such as "Hong Kong," "barricades," "occupy central" and "umbrella" were censored on Sina Weibo, a popular Twitter-like social media platform in mainland China.[6] The official line of Chinese media was to cover the protests, but focusing on blue-ribbon themes. CCTV focused on the negative consequences

[3] http://cpc.people.com.cn/n/2014/1003/c87228-25774432.html, accessed on April 16, 2016; http://www.rfa.org/mandarin/yataibaodao/gangtai/xl2-10022014102343.html, accessed on April 16, 2016.

[4] http://www.nytimes.com/2014/10/18/world/asia/china-is-directing-response-to-hong-kong-protests.html?_r=0, accessed July 24, 2015. http://www.ejinsight.com/20150326-has-leung-really-secured-beijings-blessing-to-seek-second-term/, accessed July 24, 2015.

[5] China Digital Times, http://chinadigitaltimes.net/2014/09/minitrue-delete-harmful-information-hong-kong/, accessed July 24, 2015.

[6] http://www.nytimes.com/2014/10/01/world/asia/chinese-web-censors-struggle-with-hong-kong-protest.html, accessed July 24, 2015.

of the protests on the economy and the responsibility of the protesters to end the illegal occupations. Elections and protesters' demands were framed as a foreign intervention in Chinese domestic affairs. People's Daily, the mouthpiece of the central Chinese Communist Party claimed that protesters were trained by foreign forces in order to undermine the authority of the government.[7]

Not surprisingly, then, protesters in Hong Kong predominantly used social platforms outside of the so-called "Great Chinese Firewall," platforms such as Facebook and Twitter, to spread information and mobilize support. Our analysis of Twitter therefore provides insights into the network connections formed between and among both citizens located in Hong Kong and international observers of the movement.

Twitter's Terms of Service

Twitter, its tools for data collection, and its terms for third-party data use provide an excellent opportunity to explore the implications of the right to be forgotten for social scientific research. Twitter maintains an historical archive of all tweets and associated metadata generated since its inception in 2006 to which scholars and others may gain (paid) access. However, Twitter removes any tweet from the historical archive that has been deleted from the platform for any reason. Thus, if a user deletes an individual tweet or closes an entire account, the associated data no longer appear in the archive. The same is true if Twitter suspends an account or removes

[7] http://opinion.people.com.cn/n/2014/0929/c1003-25761887.html, accessed July 24, 2015. http://www.nytimes.com/2014/09/01/world/asia/hong-kong-elections.html, accessed July 24, 2015. See also http://cmp.hku.hk/2014/10/10/36410/, accessed July 24, 2015.

spam. Even retweets are removed when the original tweet is deleted. In short, substantial amounts of historical Twitter data disappear or "decay" over time.

Moreover, as part of its terms of service agreement, Twitter requires that third parties "respect users' control and privacy" by deleting any "Content that Twitter reports as deleted or expired," as well as any content that has been changed from public to private.[8] As such, Twitter's terms of service require that researchers recognize the right of users to control access to their personal data at any point in time, regardless of whether it was once available to the public. This, in turn, places a researcher's data in a constant state of flux. With data perpetually decaying, pure reproducibility—whereby one verifies results using the same set of data and following the same analytical procedures—is by very definition impossible. And if one is interested in examining the same issue or event, it may also impact replicability—or the process of testing one study's results using similar research procedures and conditions, but employing new data. That is, the robustness of our findings may be called into question by subsequent studies relying on incomplete and potentially biased data.

The Umbrella Movement Twitter Data

Just how much might we expect the data to vary over time? And how different might the conclusions we draw from these data be? In order to answer these questions we have gathered two Twitter datasets. Both capture tweets, including retweets, sent between October 1st and October 15th, 2014 containing one or more of the following popular hashtags: #HongKong,

[8] Twitter Developer Policy, https://dev.twitter.com/overview/terms/policy, latest version effective May 18, 2015.

#OccupyCentral, #UmbrellaRevolution, #OccupyAdmiralty, #HK929, and #HKStudentStrike.[9] The first dataset was obtained by purchasing tweets from Twitter's historical archive via Sifter, one of a handful of third-party applications licensed to search, retrieve, and re-sell archive data.[10] We collected the archive data on December 21, 2014, just after the Hong Kong occupations ended. However, the fact that we obtained the data two months after their origination means that even this dataset does not represent a complete record of relevant Twitter activity. Indeed, only Twitter's so-called "Firehose" application programming interface (API) offers real-time capture of the full stream of public tweets, but, as of writing, access to the Firehose costs around $3,000 per month and requires substantial technical and infrastructural support, placing it out of reach for the vast majority of social scientists. Because we were interested in how many and which tweets had been deleted over time, we used the archive dataset as the starting point for the second round of data collection. Each tweet contains a unique id number that can be used to capture the tweet and its associated metadata from another Twitter API, the REST API, which is open to the public and free of charge. Thus, on December 30, 2015, we queried the REST API with the full list of tweet ids found in the 2014 data. Any tweets that had not been deleted as of December 30, 2015 were thereby recaptured.

The archive dataset contains 556,412 tweets, while the recapture dataset comprises 506,356 tweets, or 91.0% of the original data. This finding is in line with previous research suggesting that internet data decays by about 10% annually (SalahElDeen, 2012). Thus, it does not seem that inordinate amounts of data rapidly disappear, even when related to an inherently contentious event such as the Hong Kong umbrella movement. Under many circumstances, we

[9] The hashtag queries were not case sensitive.
[10] See http://discovertext.com/sifter/.

might be satisfied with the recapture dataset. Following the law of large numbers, and because we are working with such high-volume data, when 91% of the data remain intact, many basic descriptive measures should remain acceptably close. Take, for example, the hashtags found in each dataset. The archive contains 24,500 unique hashtags, the recaptured dataset, 23,248, or 93.0%. The average number of times each hashtag appears in the archive data is 49.88, while average frequency is 46.72 in the recaptured data. Indeed, a t-test reveals there is no difference in the means between the two datasets.

Bias in Social Network Analysis

However, few social scientists will be interested in such broad, aggregate findings alone. Indeed, a great many scholars are particularly interested in the social networks resulting from social media data, including Twitter. Communication researchers studying discourse and framing, for instance, frequently analyze hashtag co-occurrence networks. Hashtags are often used to signal topics or to otherwise express intent and meaning within a tweet. Take the six hashtags we used to collect our own Twitter data. Each conveys, at a minimum, that the tweet is associated with the events occurring in Hong Kong. However, when these hashtags appear in the same tweet with additional hashtags, we are often able to identify deeper meaning and intent. We might expect, for instance, that when #OccupyCentral co-occurs in a tweet with #Democracy, the Twitter user sees the movement as a campaign for democratic freedom and likely supports the movement’s agenda. On the other hand, a tweet that contains both #OccupyCentral and #BlueRibbon likely denies the legitimacy of, and supports the police response to, the Hong Kong protests. Thus, hashtag co-occurrence networks might be used to examine the topics discussed online during a given event, to analyze how frames spread via social media, to explore how discourse and framing shift over time, and so on. Many scholars also construct and analyze user and mentions networks in order to examine digital leadership dynamics within organizations, events, or both. Directed user-mention networks reproduce the connections formed when one Twitter user mentions another by using the latter's @username in the body of the tweet. Users who frequently mention others often serve as diffusers in a social movement network, helping to spread information to or about many others. Those who are frequently mentioned, on the other hand, are often key leaders within a movement. Similarly, mention co-occurrence networks (i.e., two mentions appear in the same tweet) may help to uncover movement or group leaders, including brokers—or those actors who serve as key links between otherwise unconnected, or at least distantly connected, actors.

Unfortunately, however, social network analyses are particularly vulnerable to biases generated by missing values. That is, when analyzing social networks, a very large sample may still produce considerable biases in our findings. In a study of digitally-gathered social networks, Wang, Xhi, McFarland, and Leskovec (2012) investigated the impact of measurement error introduced as a result of missing nodes and edges (edges are another term for the connections or ties between nodes). Randomly removing ever larger numbers of nodes and edges from these digital networks, they found that networks with positively skewed degree distributions often tolerate low levels of measurement error. In such networks, a small proportion of nodes have a large number of connections, while many have just one connection. As such, if even a small number of highly-connected nodes or, similarly, a small number of edges tied to highly-connected nodes, are removed, the relative position of all the nodes within the network can change rather significantly.

And this is precisely the type of network we are most likely to observe from Twitter data. Because, for example, only a small handful of users tweet at high volume, very few hashtags trend, and so on, most Twitter networks are likely to have positively skewed degree distributions. Thus, even if Twitter data decays somewhat slowly, just a small amount of missing data may result in significantly biased network measures.

We test this expectation empirically by analyzing several network graphs and metrics generated from our two datasets. Each graph corresponds with one of the three commonly applied networks described above: hashtag co-occurrence, mention co-occurrence, and directed user-mention networks. In the latter case, a person tweeting (the user) “directs” or “sends” a connection to each person or entity s/he mentions in a given tweet; the person mentioned therefore “receives” this connection.

For all three network types we compare the number of nodes and edges, as well as three common node-level network measures—degree, betweenness, and eigenvector centrality—across our two datasets. Degree centrality represents a count of the number of ties each node has. Thus, in the hashtag co-occurrence network, if #HongKong occurs in tweets alongside 20,000 other hashtags, its degree centrality measure is 20,000. Note that in our data these are not unique observations. If #HongKong appears with #Democracy 2,000 times, each of these co-occurrences counts as a tie (in social network terminology, we therefore have “weighted edges”). In our user-mention directed networks, we measure both in-degree centrality (i.e., the number of mentions *received*) and out-degree centrality (i.e., the number of mentions *sent*).

Our second network metric, betweenness centrality, is a measure of brokerage or gatekeeping. It measures how often a given node falls along the shortest path between two other nodes. Nodes with high betweenness centrality are typically presumed to control access and information within a network. At the very least, they have the potential to significantly disrupt the flow of information (Borgatti et al, 2013, 174).

The third metric, eigenvector centrality, is a variation on degree centrality that takes into account a node's own ties, plus the ties of its neighboring nodes, plus the ties of the neighbors' neighbors, and so on. The reasoning here is that a node is especially influential if those to which it is connected are also influential. In other words, a node that has many ties to nodes that are otherwise unconnected is much less important than a node that has many ties to nodes that are themselves highly connected. Eigenvector centrality is thus a measure of relative influence or popularity within a network (Borgatti et al, 2013, 168).

We are ultimately interested in the robustness of each of these metrics as we move from the archive to the recaptured dataset. We evaluate robustness in two key ways: first, by calculating the relative difference, or error, between the centrality score observed for a given node in the archive and recapture networks; second, by comparing the ordered rankings of nodes by each centrality measure. We describe both of these procedures and their results in detail below.

Findings

First, however, let us take a look as some of the basic network characteristics. Table 1 offers a set of descriptive statistics for all six network graphs drawn from the archive and recaptured data. Across all graphs, the vast majority of nodes and edges are present in the recaptured data. At the lowest end, 84.07% of edges were recaptured in the hashtag co-occurrence network, and 88.58% of nodes were recaptured in the mentions co-occurrence network. The table also confirms that each of the networks derived from the archive data have positively skewed degree distributions.

The positive skew is particularly high in the hashtag and user-mention graphs. We therefore expect all three networks to be highly susceptible to bias, but the hashtag and user-mention networks especially so.

Table 1: Descriptive Statistics for the Archive and Recapture Networks

		Nodes		**Edges**		**Degree Skewness**
		Count	**% Recaptured**	**Count**	**% Recaptured**	
Hashtags	**Archive**	24,500	94.89	1,494,751	84.07	117.88
	Recapture	23,248		1,256,621		114.90
Mentions	**Archive**	15,301	88.58	207,509	90.36	30.04
	Recapture	13,553		187,504		20.00
User-Mention	**Archive**	145,728	91.84	586,738	90.76	158.02
	Recapture	133,833		532,486		153.80

Network Measurement Error

The first calculation we use to assess the robustness of our social network metrics is measurement error. In statistics, measurement error is understood as the difference between the true value of an item and its observed value. Because we do not have the full population of tweets meeting our data collection criteria, we cannot assess the centrality measures generated by recaptured data against their true values. However, we are able to use the archive data as a near approximation. We therefore gauge the relative error generated in the recaptured data by calculating the difference between the centrality scores observed in each network for a given node and dividing by the node's score in the archive network. Take, for example, the degree centrality scores for #HongKong in the hashtag co-occurrence networks. In the archive network graph, this hashtag has 20,595 ties. In the recapture graph, #HongKong has 19,508 ties. Its relative error is therefore 0.0528, or ((20,595 – 19,508) / 20,595). In other words, #HongKong's degree centrality score in the recapture network is 5.28% lower than that of the archive.

To more fully illustrate these calculations, Table 2 provides a comparison of the degree, betweenness, and eigenvector centrality scores, as well as their relative errors, for the top five nodes in the archive mentions co-occurrence network. Note that the top mention in terms of both degree centrality and betweenness centrality does not appear at all in the recaptured data. For those nodes that do not appear, we assign the maximum relative error value observed across the networks. For degree centrality, the maximum value is 1.00. No node can have more connections in the recaptured graph than it does in the archive, and a completely disconnected node has degree 0. However, for betweenness and eigenvector centrality, where missing nodes and edges can substantially lower or raise the centrality of other nodes, relative error can be much higher. As Table 3—which presents the mean relative error for all nodes in the hashtags, mentions, and user-mention networks—shows, the maximum values associated with betweenness centrality are particularly high.

Table 2: Relative Error, Mentions Co-occurrence Networks

Degree Centrality			
Mention	**Score**		**Relative Error**
	Archive	**Recapture**	
rightnowio_feed	1072	--	1.0000
hkdemonow	699	645	0.0773
oclphk	572	551	0.0367
tomgrundy	426	374	0.1221
scmp_news	377	353	0.0637
Betweenness Centrality			
Mention	**Score**		**Relative Error**
	Archive	**Recapture**	
rightnowio_feed	13,976,953	--	36,625.1523
hkdemonow	5,923,100	4,920,030	0.1693
oclphk	5,636,451	5,794,951	-0.0281
hk928umbrella	3,881,635	2,989,842	0.2297
wsj	3,420,049	3,020,590	0.1168
Eigenvector Centrality			
Mention	**Score**		**Relative Error**
	Archive	**Recapture**	
hkdemonow	1.0000	1.000	0.0000
williamsjon	0.9915	0.9915	0.0000
panphil	0.1907	0.1900	0.0041
france7776	0.0739	0.0738	0.0016
kemc	0.0636	0.0637	-0.0014

Following Wang and colleagues (2012), we presume that the bias introduced by very low levels of error, 0.0500 or less, is likely to be “trivial” (407). But above this level, bias is likely to have a more substantial impact on one’s findings. As Table 3 shows, the mean relative error is above 0.0500 for all but the in-degree centrality scores found when comparing the user-mention networks. Degree centrality proves to be the most robust measure for each network category. In comparison, betweenness centrality proves exceedingly prone to error, with the mention co-occurrence networks demonstrating an average relative difference of 0.6207 and the hashtag networks a remarkable 4.2908.

Table 3: Mean Relative Error

		Mean Relative Error	Standard Deviation	Maximum
Degree Centrality	**Hashtags**	0.0665	0.2309	1.0000
	Mentions	0.1411	0.3259	1.0000
	User-Mention, In-degree	0.0210	0.1297	1.000
	User-Mention, Out-degree	0.0805	0.2669	1.0000
Betweenness Centrality	**Hashtags**	4.2908	276.4510	36,625.1523
	Mentions	0.6207	21.0771	2,207.8168
	User-Mention	0.1254	15.4421	3,398.5269
Eigenvector Centrality	**Hashtags**	0.1412	0.2166	1.0000
	Mentions	0.2000	0.3313	2.9997
	User-Mention	0.0510	0.1959	2.6848

Correlation of Centrality Rankings

These results already raise serious concerns about biases resulting from analysis of the recaptured data. However, mean relative error does not provide the whole picture. The distribution of these errors across a network also matters. Even when the mean error for all nodes is quite low, if that error is distributed unevenly—and particularly if larger errors are associated with the most central actors—it is likely to have serious consequences for our findings. When interpreting network data, one is usually particularly interested in the most central nodes. Using the hashtags data, one might focus on the most prominent hashtags and their connections in order to unpack and understand the dominant topics or discourses. Examining the mentions or user-mentions networks, our interest is likely to be in the most influential actors and their roles in the networks. But if we are misidentifying who those actors are due to measurement error, our conclusions will be fundamentally flawed.

With this in mind, the second method we use to assess the robustness of our social network metrics employs Kendall's tau correlations of rank ordered lists for each of the centrality measures. Kendall’s tau gauges the ordinal association, or the similarity of the rank orderings, between two lists. In order to illustrate some of these rank orderings as they appear in the Hong Kong data, Table 4 offers a comparison of the top 10 nodes in the mentions co-occurrence networks for each centrality measure.

To calculate Kendall’s tau, we take the archive data and the rank orderings that result from those data as baseline. As with the relative errors, we use 0.0500 as the cutoff point, presuming that correlations of 0.9500 or higher are likely to result in minimal levels of bias. Table 5 displays the correlation coefficients for each centrality measure based on lists of the top 10, 25, 50, 100, 250, 500, and 1000 nodes in each network.

In total, only 12 out of 70 (17.14%) correlation coefficients are 0.9500 or higher, and most fall substantially below this threshold. Degree centrality is again most robust, with an average correlation coefficient across all three networks of 0.8662, and it is particularly robust for in-degree rankings in the user-mention network. Indeed, this is the only metric for which the mean correlation for all lists—from top 10 to top 1000—is higher than 0.95000. On the other hand, with an average coefficient of 0.6858 across the hashtags, mentions, and user-mention networks, betweenness centrality is again least robust.

Table 4: Top 10 Nodes in the Mentions Co-Occurrence Networks

Degree Centrality					
Archive			Recapture		
Rank	Node	Score	Rank (archive)	Node	Score
1	rightnowio_feed*	1072	2	hkdemonow	645
2	hkdemonow	699	3	oclphk	551
3	oclphk	572	4	tomgrundy	374
4	tomgrundy	426	5	scmp_news	353
5	scmp_news	377	6	wsj	294
6	wsj	323	7	bbcworld	273
7	bbcworld	291	8	williamsjon	245
8	williamsjon	257	9	time	231
9	time	249	11	hk928umbrella	221
10	krislc	245	10	krislc	217
Betweenness Centrality					
Archive			Recapture		
Rank	Node	Score	Rank (archive)	Node	Score
1	rightnowio_feed*	13,976,953.11	3	oclphk	5,794,951.49
2	hkdemonow	5,923,100.21	2	hkdemonow	4,920,030.04
3	oclphk	5,636,450.54	5	wsj	3,020,590.47
4	hk928umbrella	3,881,634.82	4	hk928umbrella	2,989,841.99
5	wsj	3,420,048.63	8	scmp_news	2,772,167.87
6	tomgrundy	3,190,610.81	6	tomgrundy	2,595,443.44
7	time	2,946,749.75	7	time	2,294,809.83
8	scmp_news	2,803,754.42	9	krislc	2,235,635.71
9	krislc	1,998,297.65	10	nytimes	1,703,180.23
10	nytimes	1,954,599.23	11	bbcworld	1,703,159.88
Eigenvector Centrality					
Archive			Recapture		
Rank	Node	Score	Rank (archive)	Node	Score
1	hkdemonow	1.0000	1	hkdemonow	1.0000
2	williamsjon	0.9915	2	williamsjon	0.9915
3	panphil	0.1907	3	panphil	0.1900
4	france7776	0.0739	4	france7776	0.0738
5	kemc	0.0636	5	kemc	0.0637
6	zuki_zucchini	0.0607	6	zuki_zucchini	0.0605
7	raykwong	0.0400	7	raykwong	0.0402
8	lisahorne	0.0236	8	lisahorne	0.0241
9	paddycosgrave	0.0180	9	paddycosgrave	0.0182
10	afp	0.0148	10	afp	0.0152

*Node does not appear in the recaptured data.

Table 5: Kendall's tau correlations

Degree Centrality				
	Hashtags	**Mentions**	**User-Mention**	
			In-degree	**Out-degree**
Top 10	1.0000	0.6000	1.0000	0.5111
Top 25	0.9583	0.7893	0.9933	0.7179
Top 50	0.9289	0.7911	0.9763	0.7843
Top 100	0.9303	0.8563	0.9595	0.8619
Top 250	0.9071	0.8473	0.9133	0.8567
Top 500	0.9016	0.8610	0.9060	0.8605
Top 1000	0.8852	0.8727	0.9152	0.8705
Mean	0.9302	0.8025	0.9519	0.7804

Betweenness Centrality			
	Hashtags	**Mentions**	**User-Mention**
Top 10	1.000	0.4222	0.2000
Top 25	0.9733	0.6000	0.3733
Top 50	0.9118	0.6424	0.5673
Top 100	0.8040	0.6962	0.5875
Top 250	0.8253	0.7291	0.6739
Top 500	0.7739	0.7066	0.7133
Top 1000	0.7591	0.7046	0.7373
Mean	0.8639	0.6430	0.5504

Eigenvector Centrality			
	Hashtags	**Mentions**	**User-Mention**
Top 10	0.9111	1.0000	0.3778
Top 25	0.9333	0.9867	0.5400
Top 50	0.8173	0.9755	0.6686
Top 100	0.7515	0.9455	0.7160
Top 250	0.7723	0.9194	0.7365
Top 500	0.8375	0.8922	0.7372
Top 1000	0.8652	0.8735	0.7112
Mean	0.8412	0.9418	0.6410

Interestingly, the correlation coefficients for eigenvector centrality in the user-mention network are very low—ranging from 0.3778 to 0.7365, with an average of 0.6410. This occurs despite the fact that the average relative error for user-mention eigenvector centrality was just

0.0510. As it turns out, this discrepancy occurs precisely because the error is distributed unevenly across the network. The 10 most central nodes have an average relative error of 0.3606.

The Kendall's tau results for the hashtag networks are also surprisingly weak. Given the fact that we collected our data by querying tweets containing one or more of six hashtags, we would expect the Kendall's tau coefficients to be very high, especially in the small (i.e., top 10, top 25) lists. Though the correlations are above the 0.9500 threshold for the top 10 and top 25 nodes based on degree and betweenness centrality, the hashtags network actually proves much less robust than the mentions network for eigenvector centrality, never rising above 0.9333.

Conclusion

Taken together, these findings suggest that honoring the right to be forgotten in social media research is likely to have substantial consequences for social scientists. Should we acknowledge that, when obtained without formal consent, we have little right to maintain—and, in particular, to *share*—data once they are removed from the public domain, the inferences drawn from such "decayed" data are likely to be considerably biased.

This seems particularly true if we are using decayed data to conduct social network analysis—though the magnitude of the impact does vary by the network metric in question. As we have seen, measurement error is extremely high for betweenness centrality measures. A year after the Hong Kong protests ended, it is clear that key data regarding brokers and the links they provide between concepts and actors in our Twitter networks were lost. Moreover, this extremely high degree of error promotes flawed conclusions regarding the relative prominence of various hashtags, users, and mentions. Degree centrality, on the other hand, is the most robust network metric. The degree of error across all the networks proved relatively low, and the ranking correlations comparatively high. And yet, only the in-degree centrality measures drawn from the user-mention networks fall within generally acceptable ranges of error. The last metric, eigenvector centrality rests between the first two metrics. Relative error is much lower than that associated with betweenness, but is still substantial. Because eigenvector centrality scores are based not just on the ties of a given node itself, but on those ties, plus the ties of its neighbors, plus the ties of its neighbors' neighbors and so forth, a small amount of missing data can quickly distort our understanding of influence and popularity within a network (Wang et al, 2012, 401). The eigenvector centrality results also provide a clear portrait of the impact that the distribution of errors can have on our findings. Even when the mean error across all nodes is quite low, if larger errors are associated with the most central actors—precisely as occurs in the user-mention network—we are likely to reach flawed conclusions regarding which nodes are most central. To be sure, looking across the Kendall's tau results, it is clear that focusing on just the top 10 or 25 nodes would generally be ill-advised. However, even as we reach deeper into the data—looking all the way to the top 1000 nodes—rank correlations remain troublingly low.

Of course, these findings are based on limited data. We would ideally like to be able to compare recaptured data to the full population of tweets meeting our selection criteria. And yet, measurement error will always be present to some extent in empirical data. Even had we been able to obtain tweets in real time from Twitter's Firehose API, technical and infrastructural perturbations on both the data sending and receiving ends would result in some degree of error. Moreover, we believe that the archive data we employ in our study represent a reasonable compromise between accessibility (i.e., they were not too costly and did not require vast technical and infrastructural resources) and proximity to the population of relevant data. These parameters place such data within reach of other social scientists who might like to explore this line of questioning.

Indeed, we believe this is a line of inquiry worth further pursuit. Our analysis has provided a set of initial findings using a single case study—the 2014 Hong Kong umbrella movement—and considered the implications for a popular, but still singular, methodology—social network analysis. Nonetheless, the implications are clear: If we wish to take the right to be forgotten seriously, scholars must begin discussions about how to best protect both the rights of their research subjects and the integrity of social scientific processes. A full-speed-ahead approach to data sharing makes the former impossible, but, conversely, a total embrace of the right to be forgotten seems likely to introduce substantial bias and undercut efforts to ensure replicability in our research.

References

Borgatti, Stephen P., Martin G. Everett, and Jeffrey C. Johnson. 2013. *Analyzing Social Networks*. SAGE Publications. Kindle Edition.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The Dynamics of Protest Recruitment through an Online Network." *Scientific Reports* 1 (197): 1-6

Hanna, Alexander. 2013. "Computer-Aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18 (4): 367–88.

Harrigan, Nicholas, Palakorn Achananuparp, and Ee-Peng Lim. 2012. "Influentials, Novelty, and Social Contagion: The Viral Power of Average Friends, Close Communities, and Old News." *Social Networks* 34 (4): 470–80.

Koops, Bert-Jaap. 2011. "Forgetting Footprints, Shunning Shadows: A Critical Analysis of the 'Right to Be Forgotten' in Big Data Practice." SSRN Scholarly Paper ID 1986719. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=1986719.

SalahEldeen, Hany M. and Michael L. Nelson. 2012. "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?" In *Theory and Practice of Digital Libraries*, Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides, eds. Berlin: Springer, 125-137.

Tsesis, Alexander. 2014. "Right to Erasure: Privacy, Data Brokers, and the Indefinite Retention of Data." *Wake Forest Law Review* 49(2): 433-484.

Tremayne, Mark. 2014. "Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street." *Social Movement Studies* 13 (1): 110–26.

Wang, Dan J., Xiaolin Shi, Daniel A. McFarland, and Jure Leskovec. 2012. "Measurement Error in Network Data: A Re-Classification." *Social Networks* 34 (4): 396–409.