Global Connectivity and Multilinguals in the Twitter Network

Scott A. Hale
Oxford Internet Institute, University of Oxford
1 St Giles, Oxford OX1 3JS, UK
scott.hale@oii.ox.ac.uk

ABSTRACT
This article analyzes the global connectivity of the Twitter retweet and mentions network and the role of multilingual users engaging with content in multiple languages. The network is heavily structured by language with most mentions and retweets directed to users writing in the same language. Users writing in multiple languages are more active, authoring more tweets than monolingual users. These multilingual users play an important bridging role in the global connectivity of the network. The mean level of insularity from speakers in each language does not correlate straightforwardly with the size of the user base as predicted by previous research. Finally, the English language does play more of a bridging role than other languages, but the role played collectively by multilingual users across different languages is the largest bridging force in the network.

Author Keywords
Social Media; Information Discovery; Social Network Analysis; Information Diffusion; Cross-language; Micro-blogs; Multilingual

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces; H.5.4 Information Interfaces and Presentation (e.g. HCI): Hypertext/Hypermedia

INTRODUCTION

Given the wide differences in information available in different languages [13, 15], multilingual users of social media platforms may be able to share (or at least indicate) novel information with users of different languages and thereby help overcome in a small way traditional language divides [7]. This paper examines the role that users engaging with content in multiple languages (referred to as multilingual users) play in the Twitter network in order to better inform the design of search and friend/follower recommendation systems on social media platforms. It analyzes the global connectivity of the Twitter mentions and retweet network to assess the extent to which language structures the network and asks whether multilingual users form cross-language bridges for information exchange that provide unique connections without which the network would break apart. The study finds language is a key force structuring ties in the network, but that multilingual users are an important case to design for: over 10% of users engage with content in multiple languages and these users are more active than their monolingual counterparts. There is thus value in a single multilingual system, and language should not be used as an absolute (i.e. only ever returning search results or recommending friends in the main language of the user).

English is the most used language on Twitter, but still accounts for less than half of the messages exchanged on the network [15]. Unlike the different language editions of Wikipedia, which are only loosely connected, Twitter takes a more integrated approach to language. Users can follow users from multiple languages and have only one profile (unlike the one profile per language of LinkedIn). Understanding how users connect across languages is important in designing the search function and follower recommendation system. To what extent should language be used to restrict or prioritize results in cases where a term occurs in tweets of many languages (e.g., a company name)? Similarly, what role should language play in recommending other users one might follow? If language is used, but few results are available in the user's preferred language(s), what secondary languages might be most beneficial to include results from?

Past work in international telephony, television, and social media have often noted that language and country may be strong impediments to communication [e.g. 3, 11, 12, 14, 18, 21, 25]. Exactly what information is shared between speakers of different languages on social media and to what extent, however, remains unclear. The handful of studies looking at language and social media have found language plays a large role in structuring the hyperlink relationships between blogs [12, 14] and the follower/following relationships between Twitter users [21].

Nevertheless, the 'weightless bits' of digital communication may render barriers of language and geography less relevant online compared to the 'heavy atoms' of the physical world [26]. Online platforms must strike a balance between reinforcing within language tendencies and allowing for the

serendipitous discovery of new content from speakers of other languages on the platform. Even if users do not understand the language of the content, awareness of the content allows users to apply machine translation and/or possibly understand the content in part (e.g. embedded pictures). On Twitter, users select other users to follow and see a feed of what these users have posted when logged in. There is no facility on the site to easily direct a message only to a subset of users, which means content authored by multilingual users is sent to all their followers irrespective of language. This inability to direct content to a subset of one's network stands in contrast to the 'circles' at the center of Google+ and the lists and groups features of Facebook. This may be positive and introduce users to new content as users may be more likely to try to understand or apply machine translation to foreign-language content from friends or acquaintances than from unknown users.

Prevailing theory in social network analysis suggests individuals tend to group together with those similar to themselves, a property known as homophily and commonly expressed by the adage, "birds of a feather flock together" [10]. This leads to networks having many clusters or groups of nodes "within which network connections are dense, but between which they are sparser" [17].

These clusters result from many factors (gender, race, age, etc.) including language [12, 14]. As these clusters come to define the social circles in which information is sought and shared, each cluster comes to contain unique information [10]. Previous research on language is consistent with this view, finding very different information is available in different languages. On Wikipedia, for instance, article overlap between language editions of the encyclopedia is low [13]. On Twitter, hashtags and links to different web domains often differ across languages [15]. Given this difference in information, cross-language connections can be extremely important. Individuals who connect different clusters may be exposed to novel information that differs from the information in a single cluster.

Previous work has found language and geography structure the following–follower network on Twitter [16, 20, 21], but has not examined the bridging potential of multilingual users. The closest work to date by Eleta and Golbeck [7] used an ego-net approach, examining the follower–following networks of 73 multilingual Twitter users and found that some multilingual users could act as unconscious translators moving information from one language to another.

This paper builds upon this work in two ways. First, it focuses on the emergent network of message sharing activity. This network is formed by Twitter users mentioning other users by username, replying to tweets, and retweeting (resending/ forwarding) tweets. This higher bar for a network tie captures the dynamic, interaction patterns rather than the more static, following network on the platform. Consistent with previous work and corresponding with the low overlap in hashtag/ link domains, it is predicted that *the mentions/retweet network will have many clusters composed of a single, dominant language* (H1). Second, this paper examines the bridging role of multilingual users at a full network level rather than an ego-net level. It specifically tests the hypothesis that *multilingual Twitter users serve as bridges between different clusters in the network* (H2).

If language does structure the network but multilingual users serve as bridges between languages as predicted, then it would be useful to know the distribution of multilinguals and how they connect users across languages when designing search and friend recommendation approaches. Strongly connected languages are likely good languages to draw additional results from for search or friend recommendations when insufficient results are available in the preferred language(s) of the user.

In what languages are users more likely to cross language divides? Qualitative and survey work has suggested that users writing in less represented languages will more likely cross-language boundaries [6, 23, 24]. Importantly, survey respondents in Uzbekistan reported crossing language boundaries online even while simultaneously reporting low confidence in their foreign language skills [24]. This leads to the hypothesis that *users writing in less-represented languages will be more likely to cross language boundaries than users writing in highly-represented languages* (H3).

When users do cross languages, linguist David Crystal [5] suggests these users will engage with content and users in larger languages, particularly English. Previous studies of language connectivity online have also suggested English plays a special, bridging role connecting speakers of other languages. Herring, et al. [14] examined LiveJournal blogs and found language to be a strong factor in structuring 'friend' relationships on the site. English served as a bridging language, and "when non-English journals friend a journal in another language, that language is almost always English." Similarly, Hale [12] examining Japanese, English, and Spanish-language blog posts about the 2010 Haitain earthquake found significantly fewer links between Japanese and Spanish than either Japanese and English or Spanish and English. Both studies, however, had relatively small sample sizes and started with a small number of languages. English may not serve as strong a bridging role when more languages are included. A mapping of the Arabic blogosphere suggested French, for instance, may serve as an important bridging language in North Africa [8]. Similarly, survey work suggests Russian could serve a bridging role in the former Soviet bloc [24]. This paper will analyze the collective role of users in different languages in the Twitter mentions/retweet network to identify which languages serve more of a collective bridging role. It specifically tests if *English-language users as a whole form more bridges than users writing in other languages* (H4). The extent to which English is a natural bridging language in the network can help determine the extent to which English is a good default or fall-back language for search and friend recommendation results when little is known about the user.

Whereas H3 predicts the languages from which users will initiate cross-language activity, H4 predicts the languages to which these multilingual users will likely connect. The two hypotheses are complementary and together predict that users

from less represented languages will cross language boundaries to engage with English-language content (and perhaps to a lesser degree content in other well-represented languages). These more-represented languages may then collectively form bridges between users in multiple less-represented languages.

DATA

The data analyzed comes from the Twitter sample stream with 'spritzer' access, which gives a 1% sample of all tweets. Tweets were collected from June 11, 2011 to June 29, 2011.[1] From each tweet, the text of the tweet was extracted and analyzed with the Compact Language Detection (CLD) framework to determine the language of the tweet. The username of the author as well as any mention or retweet of another user were also extracted.

Language identification is difficult on such short text [4], but previous research has found the CLD kit to produce acceptable results for a wide range of languages on Twitter (while recognizing limitations such as not being trained to recognize romanizations of languages with another traditional script) [9]. CLD was developed by Google and is used within Google Chrome to detect the language of content. Urls, hashtags, and mentions were temporarily removed from the text of tweets for language detection following the recommendations of Graham, et al. [9]. CLD distinguishes between traditional and simplified Chinese as well as between Indonesian and Malay. However, given the similarity in these pairs, the two Chinese scripts were treated as one language (zh). Similarly Indonesian was included with Malay (ms).[2] Given the difficulties with shorter text it is useful to establish a threshold under which the detection of a language is more likely classifier error than authentic use of the language. For this study, a user was only considered to use a language if at least 20% of the user's tweets and at least two tweets were detected in that language.[3] Any user meeting this requirement for two or more languages was classified as a multilingual user, while the remaining users were classified as monolingual users. All multilingual users, therefore, authored at least four tweets in total (two tweets in each of two languages at a minimum). Users with less than four tweets were excluded entirely to avoid having any users in the sample with insufficient data to determine if they are monolingual or multilingual in their Twitter usage.

User mentions (@user) and retweets of another user's content were extracted to form weighted edges of the network. In the final network each node represents a Twitter user and each weighted, directed edge e_{ij} represents the number of tweets user i authored that mentioned or retweeted user j. Each node also has the total number of tweets in the sample authored by that user, the user's most-used language, and the percentage of the tweets by that user in the user's most-used language.[4]

Language	User count	Avg. tweets/user	(s.d.)
English (en)	375,474	8.43	(5.81)
Japanese (ja)	137,263	9.51	(8.38)
Portuguese (pt)	133,501	7.95	(5.18)
Malay / Indonesian (ms)	106,223	8.44	(5.51)
Spanish (es)	70,246	8.01	(5.18)
Dutch (nl)	31,035	8.81	(5.84)
Korean (ko)	16,123	10.46	(8.96)
Thai (th)	8,629	9.03	(6.48)
Arabic (ar)	7,679	8.30	(6.48)
French (fr)	5,769	9.06	(6.71)
Filipino / Tagalog (fil)	5,393	6.74	(3.64)
Italian (it)	4,795	9.03	(6.17)
Turkish (tr)	3,759	7.33	(4.49)
German (de)	2,299	8.15	(5.38)
Russian (ru)	2,282	7.72	(6.25)

Table 1. Languages with the most users and the average number of tweets per user. Each user is placed in the language he or she uses most frequently.

The nature of the Twitter API forces this study to exclude the least-active users. The distribution of the average number of tweets per user is heavy-tailed with most users writing only one tweet and a very small number of users writing a large number of tweets. To classify users as monolingual or multilingual, multiple tweets must be observed. Raising the number of tweets required improves the quality of language detection, but also decreases the number of users with sufficient data to remain in the sample. This problem is aggravated by working with only a 1% sample of tweets, but is mitigated in part by the longer length of data collection. Nonetheless, excluding the least-active users is not a major limitation as these users send only a small number of tweets each, and they are unlikely, therefore, to have any sustained bridging role in the wider network. In contrast, the most active users have a greater probability of being at the core of the network, and the cross-language activity at the core of the network is more likely to affect a wide number of users. Spam accounts are another challenge with Twitter data. To reduce the presence of these and focus on the role of humans in the network, the network is filtered to include only users receiving at least one mention/retweet ($indegree \geq 1$). Furthermore, only the largest weakly-connected component is selected.

This results in a network with 916,836 nodes and 2,652,618 directed edges. The corresponding undirected network with every directed edge converted to an undirected edge and mutual edges combined has 2,380,675 undirected edges (i.e. there are 271,943 mutually connected users in the directed network).

[1] This was an uneventful period with no large geopolitical events, and the data thus aims to approximate the average background level of activity rather than examining a particular event as some past studies have done.

[2] Grouping these pairs together actually reduces the magnitude of the role multilingual users play, but is more linguistically appropriate.

[3] 20% was chosen through manual examination of ad hoc subsets of multilingual users, which found the most prolific users had slightly less than 20% of their tweets misclassified as the wrong language in the worse cases. The 20% cutoff affects only 6% of all users. Without this bar all the results hold, often with larger effect sizes making multilingual users appear even more active and more responsible for the bridges in the network.

[4] The code used to record the stream (PHP), construct the network (Java / Hadoop), and perform the analysis (R / igraph) are available at `http://www.scotthale.net/pubs/?chi2014`

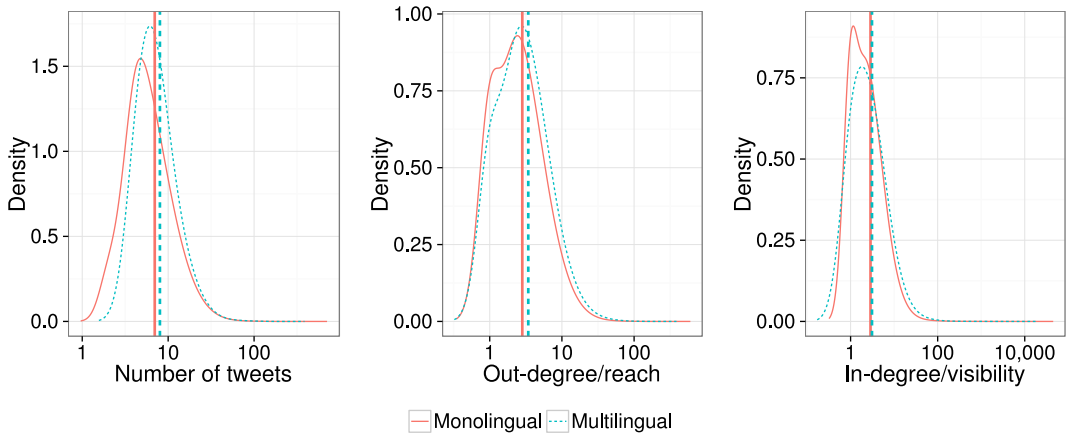

Figure 1. Density plots comparing tweet count, out-degree, and in-degree for multilingual and monolingual users. Vertical lines show mean values. All differences are significant beyond the 99.99% level as established by t-tests on the sample means.

ANALYSIS

This section first compares multilingual and monolingual users. It then investigates the possible bridging role multilingual users play before looking at the isolation/insularity of speakers of each language. It finishes with an analysis of the bridging role played by speakers of different languages and the specific role of English in the network. After this, the paper discusses the results of the study and suggests areas for further research.

Of the over 916,000 Twitter users in the sample, 103,645 (11%) were observed to use more than one language and designated as multilingual users. The distribution of messages sent per user is highly skewed, reflecting a heavy-tailed distribution among both monolingual and multilingual users. A list of the languages with the most users is given in Table 1. This list is broadly in line with the previous work by Hong, Convertino, and Chi [15] even though a different language classification approach was used (LingPipe and Google's Language API vs. the compact language detection kit). The list reflects varying uptake of the Twitter platform across the world. The conspicuous absence of Chinese, a very large language on the web, likely reflects the difficulty of accessing Twitter in mainland China as well as the availability and popularity of homegrown alternatives.

Multilingual users, on average, were more active than their monolingual counterparts sending a mean 8.0 (median 7, sd 5.2) messages per user on average compared to a mean 7.0 (median 5, sd 5.6) messages per user among monolinguals. Multilinguals also have a higher out-degree (mean 3.4 vs. 2.8; medians both 2, sd 3.9 and 3.3 respectively), and a slightly higher in-degree (mean 3.1 vs. 2.9; medians both 2, sd 14.1 and 21.5). Figure 1 shows these comparisons.

Language and network structure

Many classic network clustering algorithms use betweenness to remove edges with the highest scores from a network until it breaks into unconnected components, and the number of components into which the network is optimally divided is usually measured with modularity, which is a goodness-of-fit measure comparing the density of edges within and between clusters [17]. These methods, however, are computationally expensive and difficult to run on a network of this size. Raghavan, Albert, and Kumara [19] developed a label propagation approach to detect clusters from the structure of the network alone. Each node is given a unique label and at each step changes its label to the label that most of its neighbors have at that time. The method does not consider edge direction or edge weight. The algorithm operates in near linear time and the authors claim that after five iterations most node labels are fixed regardless of the network's size. To achieve the best performance, the author wrote a custom implementation of this algorithm to parallelize it and make use of multiple processors/cores on modern computers.[5]

After 72 runs, 99.96% of the nodes had fixed their labels. Among the 17,480 clusters found by the algorithm, 12,740 (72.9%) of the clusters were formed of users who all shared one common most-used language. This is significantly higher than the 22 clusters (0.12%) that would be expected to share a common language if language assignment was independent of network structure.[6] Indeed, the number of languages in all clusters (mean 1.4, sd 1.1, median 1) was significantly lower than what would be expected randomly (mean 4.2, sd 2.4, median 3.7). Many of the clusters found were relatively small (mean size of 52, median size of 6), while a few were extremely large (see Figure 2).

Seven clusters had more than 10,000 users each, and collectively held 61% of all the users in the graph. Four of these seven clusters are heavily dominated by speakers of one language as shown in Table 2 and Figure 3. English users[7]

[5]Code at `http://www.scotthale.net/pubs/?chi2014`

[6]Random expected values were formed by keeping the existing network and clusters, but shuffling/permuting language labels and assigning them to nodes randomly. This was done 100 times and the results averaged.

[7]An unfortunate limitation of the English language is that it often uses the same adjective to refer to both the people of a given country and people who speak the language of that country. That is, "English users" can refer to either users from England or English-language speaking users. In this paper, all such terms refer strictly to languages and never to countries/locations.

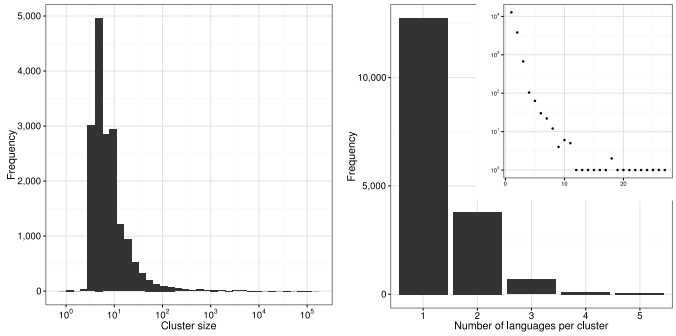

Figure 2. Histograms of the size of clusters found by the label propagation algorithm (left) and the number of languages within the clusters (right). The right histogram is truncated, while the insert displays the full distribution on a log-log scale.

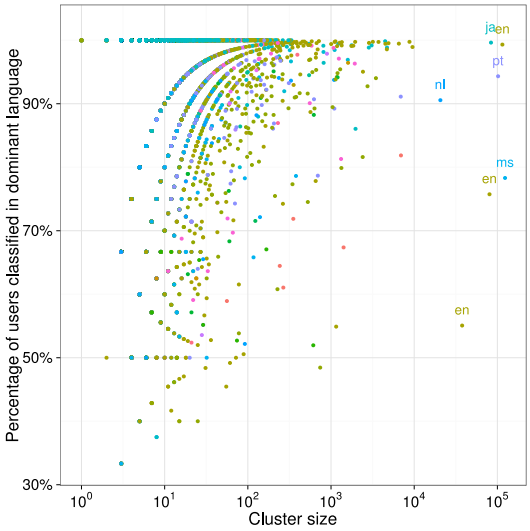

Figure 3. Scatter plot of cluster size and the percentage of users in the cluster most often using the most prevalent language. Details of the largest clusters are given in Table 2.

are the most numerous language group in three of the seven largest clusters. Ninety-nine percent of the 114,826 users in the second-largest cluster use English most frequently. In contrast, however, the other two English-language clusters while smaller are less dominated by English: only 75% and 55% of the users in these clusters use English most frequently, suggesting users in these clusters might be more cross-lingual in their communications.

The mentions/retweet network is more dynamic, representing only active communication patterns (compared to the more passive, static follower–following network), but it is also more volatile. Some of the smaller sized clusters found may have grouped together given a larger or longer data sample. Even so, most of these clusters would likely remain dominated by one language as most of the largest observed clusters are.

As mentioned previously, modularity is a measure of the goodness-of-fit of a given division of a network. A value of zero indicates the division is no better than random. (Theoretically the value has a lower limit of -1 with negative values indicating worse than random divisions, but such values are rarely observed.) Values approaching the maximum value, 1.0, indicate the network divides easily into densely connected clusters with sparse connections between them. Newman and Girvan [17] state typical values are usually in a range from 0.3 to 0.7. The divisions found through the label propagation algorithm have an extremely high modularity score of 0.81, indicating network is highly clustered. Simply dividing the network into groups *a priori* based on the majority language of users (i.e. all English users in one group, all German users in another, etc.) results in a modularity score of 0.68. Although not as strong a division as the groups found by the label propagation algorithm, the modularity score still indicates dividing the network by language alone captures much of the clustered structure in the network.

Most-used language	% users in most-used language	Number of languages	Number of nodes
Malay (ms)	78.3	41	123,616
English (en)	99.3	39	114,826
Portuguese (pt)	94.3	40	101,987
Japanese (ja)	99.6	19	83,785
English (en)	75.7	44	80,387
English (en)	55.1	42	37,688
Dutch (nl)	90.6	23	20,634

Table 2. Clusters with over 10,000 nodes found through the label propagation algorithm.

The structure found with the label propagation algorithm represents a strong division of the network, and the analysis shows that far more of the clusters found are composed of a single, dominant language than would be expected randomly, confirming H1. This claim is further strengthened by the high modularity score for a deductive division of the network based solely on language.

Bridging role of multilinguals

Do multilingual users form unique bridges connecting different clusters in the network? Depending on the fragility of the network and the position of users, removing a user may disconnect a large portion of users from the largest connected component. If multilingual Twitter users are more often than random situated in bridging positions, then removing multilingual users from the network will result in a smaller largest connected component than removing the same number of users randomly. Similarly, removing users in bridging positions should result in more disconnected components, and these components should be larger (i.e. more than isolates).

Figure 4 removes four different subsets of users from the network and records the size of the resulting largest, weakly-connected component, the number of components created, and the average size of these components (excluding the largest connected component). In the *Multilinguals* condition, all 103,645 multilingual users are removed from the network. In the *Monolinguals* condition, an equivalent number of monolinguals are chosen randomly and removed from the

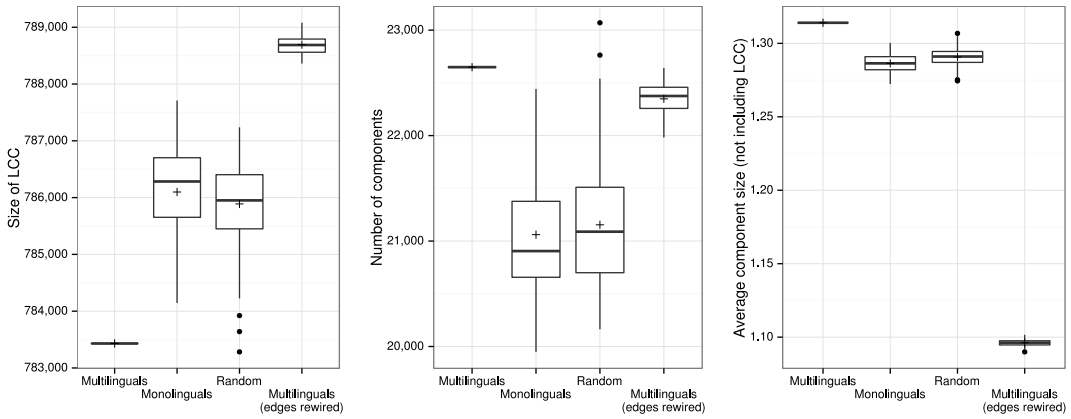

Figure 4. Size of the largest, weakly-connected component (left), total number of components (center), and average size of the components (right) created by removing all multilingual users, an equivalent number of monolingual users randomly, an equivalent number of all users randomly, and removing all multilingual users from a network with the same degree distribution but with edges randomly shuffled. Box plots show values from 100 realizations. Mean values are indicated with +.

network. In the *Random* condition, an equivalent number of users are chosen randomly without respect to their monolingual or multilingual classification (i.e. both monolinguals and multilinguals are potentially selected) and removed from the network.

Finally, the graphs also show a fourth condition, *Multilinguals (edges rewired)*, which provides a more stringent comparison to random. To ensure it is not just the higher degree of multilingual users responsible for a greater number of components, the statistics are calculated on a graph with the same degree distribution but with random edge wirings. This graph is formed according to the algorithm developed in Viger and Latapy [22]. The multilingual users removed from this graph have exactly the same (undirected) degree distribution as the multilingual users in the empirical network, but connect different nodes. Similarly, the remaining monolingual users have the same degree distribution as the empirical network. One-hundred realizations of the random selection of nodes are performed for all conditions, and the distributions of the results are plotted in Figure 4.

All three comparisons suggest multilingual users are more often in unique bridging positions than monolingual users and than random. Removing multilingual users results in a significantly smaller largest connected component than removing the equivalent number of monolingual users or an equivalent number of users randomly. Removing multilingual users also results in a larger number of components even in comparison to the edge rewiring condition which controls for the difference in degree. Moreover, these components are on average larger than in any other condition.

This analysis confirms H2 and shows that multilingual users play a unique bridging role in the network and are critical to the global connectivity of the network.

Variations by language

There is variation in the percentage of users classified in each language as multilingual. Only 1% of users writing primarily in Japanese and 2% of users writing primarily in Korean also

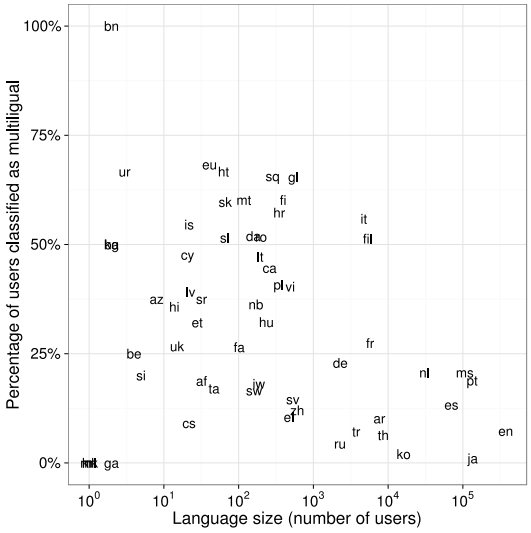

Figure 5. Number of users in each language compared to the percentage of these users classified as multilingual.

wrote in another language. On the other hand, over half of the users writing primarily in Tagalog/Filipino (fil) or in Italian (it) also wrote tweets in another language. These are the only two languages with over 1,000 users in the sample with a high level of multilingualism: see Figure 5, which shows a scatter plot of the logarithm of the number of users primarily using a language and the percentage of those users detected to also use another language. Contrary to the predictions that smaller languages would have more multilingual users, there is large variation in the percentage of multilingual users in languages with less than 1,000 primary users. Overall, there is only a weak correlation (-0.25) between the log of language size and the percentage of multilingual users.

Examining the languages of the users mentioned or retweeted reflects how much attention each user gives to other users within his/her language compared to users outside of his/her language. Averaged across all users of each language, this

Language	Language size (%)	Within language edges (%)
ja	14.97	99.49
ko	1.76	98.70
tr	0.41	95.59
th	0.94	94.33
ru	0.25	93.94
pt	14.56	93.10
nl	3.39	91.54
en	40.95	90.32
ar	0.84	90.22
es	7.66	89.25
fr	0.63	87.54
de	0.25	85.87
ms	11.59	80.48
it	0.52	54.41
fil	0.59	21.34

Table 3. Language size and measure of insularity.

gives a measure of the collective isolation or insularity for users in each language. The expected percentage of links between users of the same language varies with the number of users writing in that language. If users mentioned users without regard for their language, the percentage of all outgoing links from users of that language that pointed to other users in the same language would be proportional to the number of users writing in the language. For example, English represents approximately 40% of all users. If the destination of each edge originating from an English-speaking user were randomly chosen about 40% of the edges would terminate at another English-speaking user.

This analysis proceeds using the directed version of the network, but looks only at languages with at least 1,000 users identified in the language. This corresponds with a large drop in the number of users in each language between Russian, with nearly 2,300 users, and the next-most used language, Chinese, with about 600 users. As in the previous figure, languages with less than 1,000 users in the sample show great variability and are likely less representative of all Twitter users using that language.

H3 predicted the insularity of users would be proportional to the number of users writing in each language. Table 3 shows the fraction of all users represented by each language and the fraction of outgoing edges from users in each language that connect to a user of the same language. If the fraction of edges connecting users of the same language were proportional to the number of users writing in each language, the second and third columns in the table would vary similarly. Instead, a very different pattern emerges. For users in most well-represented languages, the number of outgoing edges connecting users of the same language is over 80% irrespective of the size of the language. Users in all languages are much more likely to mention/retweet users of the same language over users of different languages. Even so, users in two languages (Fillipino and Italian) do mention or retweet users writing in other languages more than users in most other languages.

Source language	Target language	Percent error
ja	ko	1,474%
th	ko	1,091%
pt	it	1,068%
ms	fil	784%
ko	th	550%
ja	th	536%
es	it	529%
en	fil	494%
tr	de	365%
en	it	350%

Table 4. Top language pairs with more than the expected number of edges.

Bridging languages

To determine which language pairs are connected more than random, it is necessary to first compute the expected number of edges from a given language to another language. The graph is collapsed so that there is one node per language. Each directed edge $i \rightarrow j$ captures the total number of mentions/retweets by users in language i of users in language j. Given the high level of insularity for each language, the analysis proceeds by only looking at edges between users of two different languages ($i \neq j$). If edge destination were simply random, each language group would receive links in proportion to the number of users writing in that language. This value is treated as the expected value and the percent error is calculated as $\frac{observed-expected}{expected}$ to compare the differences between these expected values and the observed values, which cover several orders of magnitude. The percent error is less than zero if fewer than the expected number of edges connect two languages, and it is greater than zero if more than the expected number of edges are found between two languages. Table 4 gives the top ten pairs of languages with more mentions/retweets than expected. There are 24 directed language pairs with no connections between them at all, and for seven language pairs there is no edge is either direction. These are indicated in Table 5, with mutually absent edges in *italics*. A force-directed graph with one node per language group and edges weighted according to the percent error is shown in Figure 6.

The network diagram and Table 4 show a small Asian-language cluster with Korean, Japanese, and Thai users more tightly connected than expected. They also show Italian and Filipino users more mentioned than the number of speakers would suggest. This corresponds with the higher level of multilingualism found for users writing in these two languages.

Similar to the analysis examining the bridging role of multilingual users as a whole, it is also possible to remove users from only one language at a time. In Figure 7, users who use each language most frequently are removed and the number of components created is shown. Users are removed in order by the percentage of their tweets in the given language. That is, users who use the language exclusively are removed first, followed by users who use the language less frequently in decreasing order. Ties are broken randomly and the lines

Source language	Target language(s)
ar	de, pt, *ru*
de	*ko*, ru, *th*
fil	de, ru
ja	*tr*
ko	ar, *de*, *nl*, ru
nl	*ko*, *th*
ru	*ar*, nl, th, tr
th	*de*, *nl*, *tr*
tr	*ja*, *th*

Table 5. Language pairs with no connections. Mutually absent pairs indicated in *italics*.

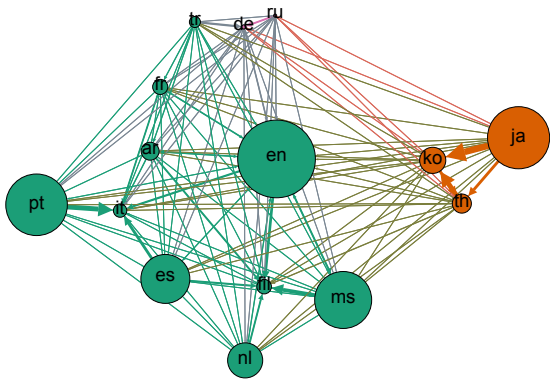

Figure 6. A collapsed network graph with users grouped to nodes representing the primary language used. Edges are weighted by the percent error in the expected vs. the actual number of mentions and retweets between language groups. Node size is proportional to the number of users primarily using each language, and node color is the result of a modularity-maximizing community detection algorithm.

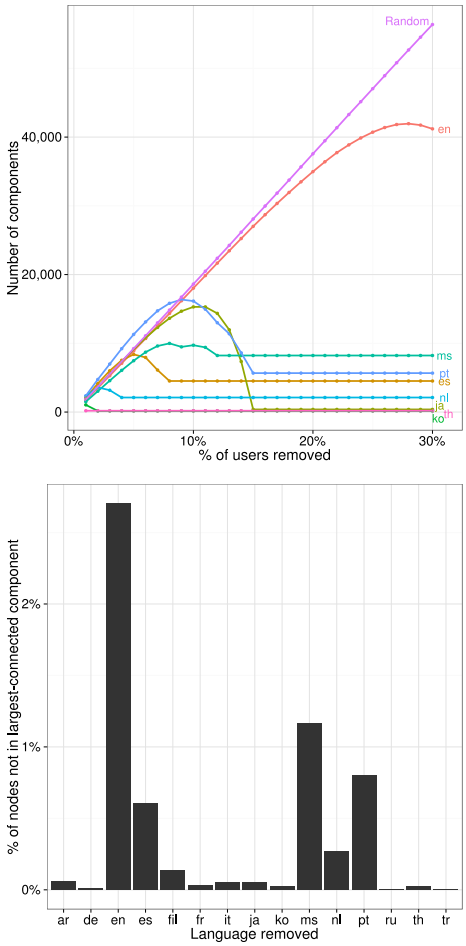

Figure 7. Number of components (top) and the percentage of remaining users not in the largest connected component (bottom) created by removing users in different languages.

show the average of 100 realizations. For each curve, users writing in a different language more frequently are left untouched. Removing users from most languages did not create more components than removing users at random.

Most curves follow a general pattern. The curves rise and reach a maximum beyond which they begin to fall. The fall in the curves corresponds to the elimination of components made entirely of speakers of the language being removed. Each curve then stabilizes on a value for the number of components created by removing all users from the given language. (The graph omits the fall of the English curve, which falls and then plateaus at approximately 11,300 components.)

Three interesting exceptions to this general pattern emerge. The lines corresponding to the removal of Japanese, Thai, and Korean users drop almost immediately reflecting a high level of insularity and few internal divisions within each language. The line corresponding to Malay users has almost no decline before plateauing. This suggests that the components created by the removal of Malay users almost always contain at least one user of another language. Finally, the line corresponding to the removal of Portuguese users initially produces more components than removing an equivalent number of users randomly. Like other languages, most of these additional components are made entirely of Portuguese users and the number of components ultimately decreases as further users are removed. The initial rise beyond the random curve suggests there are a number of subdivisions within Portuguese users themselves (possibly a Brazil–Portugal split, but this is not specifically tested).

It is clear that English serves the largest bridging role in an absolute sense, but it is also the language with the largest number of users. To calculate the relative extent of bridges created by speakers in a given language, all users from a given language are removed from the graph. Then the percentage of remaining other-language speakers not in the largest connected component is calculated. If speakers of a language do not collectively serve much of a unique bridging role, then almost no nodes will be separated from the largest connected component. Conversely, removing users from a language that collectively serves a large bridging role will leave a larger percentage of users separated from the largest connected component. The bottom graph of Figure 7 shows these percentages after removing all users from a given language. Although all the values are relatively small, removing English-language users also leaves a larger percentage of users disconnected from the largest connected component than removing users

from any other language. Users writing in English, therefore, play more of a bridging role than users writing in other languages, confirming H4.

Removing Portuguese users leaves a relatively high percentage of users disconnected, but this is somewhat expected given the large number of users writing in Portuguese. This is contrasted by the case of Japanese, however, which again reveals the insularity of Japanese speakers in the data. Despite the high number of Japanese speakers in the sample (second in size only to English) completely removing all of these Japanese users has minimal effect on the connectivity of the network.

Overall, removing all the users from any one language never disconnects as many nodes as removing all multilingual users. English users, for instance, represent three and a half times as many users as the total number of multilingual users from all languages; yet, removing this smaller number of multilingual users disconnects more nodes from the largest connected component than removing all English users (3.7% vs. 2.7%). Thus, no one particular language (not even English) appears to create as many bridges as multilinguals from multiple languages. It is the communication dynamics of multilinguals from many languages that create the network bridges found earlier.

DISCUSSION

The analysis of this sample reveals the important role played by multilingual users of Twitter. The network of retweets and mentions is heavily structured by language with most users retweeting and mentioning only users authoring content in the same language as themselves (H1). Nonetheless, users authoring content in multiple languages play a unique bridging role that is not duplicated by other users in the network (H2). This suggests that language is likely a useful feature to be used in search and friend recommendation algorithms. Even so, a balance must be struck as there are clusters of users spanning different languages. In addition, over 10% of the users in the sample used multiple languages. Any use of language, therefore, should allow for each user to have a set of multiple preferred languages and not restrict the user to one. This is especially important as multilingual users are more active than their monolingual counterparts authoring more tweets and replying/mentioning more users.

In using language for search results, it would be useful to infer a set of possible secondary languages to draw results from when there are insufficient results in the primary languages. Although there are some clear geo-linguistic patterns in the language–language network (e.g., an Asian language grouping), the level of overall multilingualism for users in each languages does not vary straightforwardly with the number of users in each language as predicted by H3. Users from less-represented languages may discuss and link to information originally from other-language sources outside of Twitter, but simply do so in isolation from users of other languages on the platform. Equally possible, these users may use a non-native language more frequently than their native tongue and thus be classified by the methods used in this paper as primary users of a different, larger language. Further research will need to look at the content and links authored by users in less-represented languages and the distribution of languages used by these users to better separate these possibilities.

Users writing most frequently in the top two represented languages, English and Japanese, play vastly different roles in the global connectivity of the network. Users in most languages direct most of their tweets to other users writing in the same language. English users are no exception, but still direct about 10% of their mentions/retweets to users in other languages. Japanese users, on the other hand, mention and retweet users from other languages only about 0.5% of the time. Of the mentions/retweets Japanese users do direct to other-language users, a disproportional number go to Korean users likely due to the geographic proximity of Japan and Korea. This indicates the importance given to same language results should be weighted differently across language groups and probably even across users primarily using the same language.

It is the unique combination of multilinguals across many language pairs that results in the global connectivity of the network. Removing speakers from one language at a time never resulted in more components than removing an equivalent number of users randomly from the network. Nonetheless, in comparison to other languages, English speaking users collectively do play a much larger role in bridging speakers of different languages than do speakers of any other language. This confirms H4, which predicted this role for English given its global prevalence and the number of second-language speakers.

While this study shows multilinguals occupy a unique place in the global retweet and mentions network, it does not examine whether foreign language content is propagated further by the monolingual followers of a multilingual user. Further research will be needed to determine whether content sent by multilingual users is further propagated by their followers in different languages. Although the Twitter platform does not offer the ability to selectively filter friends' tweets by language, users may mentally filter these tweets and stop their propagation by not engaging with them. Future research should examine specific diffusion paths (e.g. diffusion cascades as in Bakshy, et al. [1]) with reference to the languages of the users involved and the content of the tweets themselves.

Although much future work remains, this study is the first to examine the global connectivity of the Twitter retweet and mentions network with reference to language, and to establish the unique role that multilingual users play in this connectivity. While the amount of cross-language bridging is small, its presence indicates there is value in one large, multilingual system over several separate, non-connected systems for each language. Experimental work could shed light on design features that could better enable multilingual users to bridge users of different languages in the network and further enhance the unique value of the single, multilingual system. In addition, systems similar to Omnipedia for viewing content across different Wikipedia language editions [2] could better enable monolingual users to directly discover interesting content in other languages on Twitter.

ACKNOWLEDGMENTS

I would like to thank Drs. Taha Yasseri, Eric T. Meyer, Sandra Gonzalez-Bailon, Jonathan Bright, Mike Thelwall, and Irene Eleta as well as the anonymous CHI reviewers who provided helpful comments on previous versions of this article.

REFERENCES

1. Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, ACM (New York, NY, USA, 2011), 65–74.
2. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 1075–1084.
3. Barnett, G. A., and Choi, Y. Physical distance and language as determinants of the international telecommunications network. *International Political Science Review 16*, 3 (1995), 249–265.
4. Carter, S., Tsagkias, M., and Weerkamp, W. Semi-supervised priors for microblog language identification. In *Dutch-Belgian Information Retrieval Workshop (DIR 2011)* (2011).
5. Crystal, D. *English as a Global Language*, 2nd ed. Cambridge University Press, Cambridge, 2003.
6. Durham, M. Language choice on a Swiss mailing list. *Journal of Computer-Mediated Communication 9*, 1 (2003).
7. Eleta, I., and Golbeck, J. Bridging languages in social networks: How multilingual users of Twitter connect language communities. *Proceedings of the American Society for Information Science and Technology 49*, 1 (2012), 1–4.
8. Etling, B., Kelly, J., Faris, R., and Palfrey, J. Mapping the Arabic blogosphere: Politics and dissent online. *New Media & Society 12*, 8 (2010), 1225–1243.
9. Graham, M., Hale, S. A., and Gaffney, D. Where in the world are you? Geolocation and language identification in Twitter. *Professional Geographer* (2013).
10. Granovetter, M. The strength of weak ties. *The American Journal of Sociology 78*, 6 (1973), 1360–1380.
11. Hale, S. A. Impact of platform design on cross-language information exchange. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems Extended Abstracts*, CHI EA '12, ACM (New York, NY, USA, 2012), 1363–1368.
12. Hale, S. A. Net increase? Cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication 17*, 2 (2012), 135–151.
13. Hecht, B., and Gergle, D. The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, ACM (New York, NY, USA, 2010), 291–300.
14. Herring, S. C., Paolillo, J. C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L. A., and Clark, B. Language networks on LiveJournal. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, IEEE Computer Society (Washington, DC, USA, 2007).
15. Hong, L., Convertino, G., and Chi, E. Language matters in Twitter: A large scale study. In *International AAAI Conference on Weblogs and Social Media* (2011), 518–521.
16. Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, K. P. Geographic dissection of the Twitter network. In *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM-2012)*, The AAAI Press (Dublin, Ireland, 2012).
17. Newman, M. E. J., and Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E 69*, 2 (2004).
18. Nordenstreng, K., and Varis, T. Television traffic: A one-way street? A survey and analysis of the international flow of television programme material. *Reports and Papers on Mass Communication*, 70 (1974).
19. Raghavan, U. N., Albert, R., and Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E 76*, 3 (Sept. 2007), 36106.
20. State, B., Park, P., Weber, I., Mejova, Y., and Macy, M. The mesh of civilizations and international email flows. `http://arxiv.org/abs/1303.0045`, 2013.
21. Takhteyev, Y., Gruzd, A., and Wellman, B. Geography of Twitter networks. *Social Networks* (2011), 1–26.
22. Viger, F., and Latapy, M. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *Computing and Combinatorics*, L. Wang, Ed., vol. 3595 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, 440–449.
23. Warschauer, M., Said, G. R. E., and Zohry, A. Language choice online: Globalization and identity in Egypt. *Journal of Computer-Mediated Communication 7*, 4 (2002).
24. Wei, C. Y., and Kolko, B. E. Resistance to globalization: Language and Internet diffusion patterns in Uzbekistan. *New Review of Hypermedia and Multimedia 11*, 2 (2005), 205–220.
25. Wilkinson, D., and Thelwall, M. Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology 63*, 8 (2012), 1631–1646.
26. Zuckerman, E. *Rewire: Digital Cosmopolitans in the Age of Connection*. W. W. Norton & Company, London, 2013.