Making social networks more human: A topological approach

Jonathan W. Berry1 | Cynthia A. Phillips1 | Jared Saia2

1Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico
2Department of Computer Science, University of New Mexico, Albuquerque, New Mexico

Abstract
A key problem in social network analysis is to identify non-human interactions. State-of-the-art bot-detection systems like Botometer train machine-learning models on user-specific data. Unfortunately, these methods do not work on data sets in which only topological information is available. In this paper, we propose a new, purely topological approach. Our method removes edges that connect nodes exhibiting strong evidence of non-human activity from publicly available electronic-social-network datasets, including, for example, those in the Stanford Network Analysis Project repository (SNAP). Our methodology is inspired by classic work in evolutionary psychology by Dunbar that posits upper bounds on the total strength of the set of social connections in which a single human can be engaged. We model edge strength with Easley and Kleinberg’s topological estimate; label nodes as “violators” if the sum of these edge strengths exceeds a Dunbar-inspired bound; and then remove the violator-to-violator edges. We run our algorithm on multiple social networks and show that our Dunbar-inspired bound appears to hold for social networks, but not for nonsocial networks. Our cleaning process classifies 0.04% of the nodes of the Twitter-2010 followers graph are violators, and we find that more than 80% of these violator nodes have Botometer scores of 0.5 or greater. Furthermore, after we remove the roughly 15 million violator-violator edges from the 1.2-billion-edge Twitter-2010 follower graph, 34% of the violator nodes experience a factor-of-two decrease in PageRank in PageRank. PageRank is a key component of many graph algorithms such as node/edge ranking and graph sparsification. Thus, this artificial inflation would bias algorithmic output, and result in some incorrect decisions based on this output.

KEYWORDS
bot detection, data sets, Dunbar number, edge strength, electronic social networks, social network analysis

1 INTRODUCTION
Researchers need human networks for social analysis and algorithmic testing. These networks would ideally have several key properties: (a) nodes are humans, (b) edges plausibly represent a social bond, (c) the network is large, that is, at least 1 million edges, and (d) the network is reasonably complete (we thus explicitly exclude local subgraphs such as ego networks). Almost none of the networks in repositories such as the Stanford Network Analysis Project (SNAP) [25] and the Laboratory for Web Algorithmics [7] have all of these properties.

For user privacy, data sources such as SNAP typically provide networks that have been stripped of any information beyond the network topology. Therefore we assume that we have only topology (nodes and edges). We explain why these networks generally are not suitable as test sets for many human-network algorithms as provided, and give a method for
making them more suitable. We show results from our method on 15 networks with disparate edge generation mechanisms.

As a secondary benefit, the new structural properties we describe might lead to better algorithms for social networks. Because social networks can have billions of nodes and edges, algorithms should exploit structure of these networks to run faster or better than is possible on general graphs. Algorithm designers might find ways to explicitly use this structure in algorithms, or in the analysis of performance for such algorithms. This could lead to provably better performance for important network-analysis primitives on social-network graphs.

1.1 | Structure of social networks and human limits

Our method for cleaning some nonhuman behavior from social networks is based on Robin Dunbar’s work studying human social networks and the strength of human ties. Although people can know up to about 1500 people to varying degrees (see Section 7), Dunbar [15] posited that a “natural” human community has a limit of about 150 people. This number, known as the Dunbar number, comes from cognitive limits and time requirements to maintain relationships. Dunbar showed, through psychological experiments, that using electronic social networks does not significantly change the number or strength of real social relationships [16]. Emotional connection still requires personal attention, and thus time. This limits the number of strong social connections a person can maintain.

Our method requires a way to estimate the strength of a human relationship using only network topology. Easley and Kleinberg [13] use a simplified binary notion of edge strength, where each edge is strong or weak. They argue that one mechanism for edge formation in a social network is the triadic closure property, illustrated in Figure 1A. If node v has strong links to both node u and node w, then nodes u and w are more likely to be connected, at least weakly, than a random pair of nodes in a network. A wedge is a path of length two: u-v-w, that is, a pair of neighbors for the middle node v. Triadic closure involves closing wedges to triangles.

We posit that the converse of the triadic closure property is also true. That is, we posit that if one or more of the links (v, u) and (v, w) are weak, then the probability that nodes u and w will become connected through their relationship with node v decreases significantly, approaching random if both are weak. See Section 7 for justification of this conjecture based on how the central node facilitates triadic closure.

Thus we assume that if a node has bounded resources for such facilitation, and devotes these resources mostly to pairs of strong links, then as degree increases, the probability of facilitating a relationship between weak pairs decreases approximately quadratically. Hence, we do not expect a real human to be involved in too many strong relationships that are part of large, dense subgraphs.

1.2 | Nonhuman behavior in electronic social networks

Electronic social networks such as Twitter are open for both individuals and businesses to join. Twitter enables automated message sending and other behavior via their API, so it is not surprising that nonhuman behavior such as that from businesses is common. Real estate companies that tweet new listings provide a service their followers want. However, there is also substantial economic incentive to manipulate connections. Having more followers can increase a business’ ranking in search results. In 2014, businesses could buy fake followers to improve their search ranking for a penny each [19].

In some important applications, the network is not open for public enrollment. The network maintainer adds human-only nodes and observed relationships to the graph. For example, law enforcement agencies working on counter-terrorism gather data on people.

Our goal is to clean enough nonhuman behavior from electronic social networks to test algorithms designed for human-only networks. Unlike electronic social network owners, such as Facebook and Twitter, we do not have access to details of the user accounts. Thus we must conservatively clean networks based on topology only. We wish to remove only those subnetworks exhibiting substantially nonhuman behavior.

1.3 | Results summary

Our primary contributions are:

- We propose a vertex attribute called the strength index and a simple model of an upper bound on this index.
We propose a network cleaning process inspired by Dunbar’s work and provide informal validation of this process on the Twitter-2010 followers’ network using ground truth from the literature, a modern tool for bot scoring, and a structural test for a particular kind of artificial subnetwork construction.

We demonstrate the impact of this nonhuman activity on PageRank and clustering coefficients.

Unlike efforts such as Botometer [14], our goal is not to classify individual nodes, or even to locate them in a spectrum of botness. The activity we detect and remove is group activity. However, the Botometer scoring procedure and some ground truth classification of Twitter nodes by Texas A&M [23] will prove useful when we validate our techniques in Section 4.

We provide code and instructions to reproduce our plots. We also provide the processed (cleaner) data sets.1

Kleinberg [13] define edge strength to be

\[ s(e = (u, v)) = \frac{t_e}{|N_u \cup N_v| - 2}. \]  

This measures the fraction of shared nodes in the neighborhoods of nodes \( u \) and \( v \). We use this method to assign edge weights to unweighted graphs.

Computing strength indices and generalized Dunbar numbers is a straightforward application of triangle enumeration, which runs efficiently on social networks [6].

We now define a new vertex property motivated by the academic notion of \( h \)-index. A researcher’s \( h \)-index is the maximum \( h \) such that (s)he has authored at least \( h \) papers that each have at least \( h \) citations. This defines a set of papers that represents the author’s most influential work. It also brings to mind the Dunbar number since in practice the \( h \)-index is bounded by a constant. The most senior leaders of a field may have an \( h \)-index of roughly 100, but it does not seem humanly possible to have an \( h \)-index of 1000.

Consider the \( d_v \) edges incident on a vertex \( v \), sorted by decreasing edge strength (Equation (1)). This is depicted notionally in Figure 2. It is notional because the figure shows a continuous curve, while the edge strengths are actually discrete values. Let \( i \) be an index with range \( \{1, \ldots, d_v\} \). Let \( r_i \) be \( \frac{i}{d_v} \), which is the fractional rank of the \( i \)th neighbor in the sorted list. This normalizes the range of \( r_i \) to (0, 1]. For example, if \( d_v = 10 \), then the strongest edge has fractional rank 1/10 and the fifth-strongest edge has fractional rank 1/2, or halfway through the neighbor list. Let \( s_i \) be the edge strength of the \( i \)th incident edge in this order.

---

1We plan to make to code available at http://cs.unm.edu/CHANT.html.

---

**FIGURE 2** The strength index. For a given vertex \( v \), sort all edges adjacent to \( v \) by strength. (A) Plot the strength vs the edge’s fractional rank. The \( i \)th strongest neighbor (edge with the \( i \)th highest strength) has fractional rank \( i/d_v \), where \( d_v \) is the degree of vertex \( v \). We truncate rectangles such as the one shown here to the shorter side, find the one with the largest area, and use it to bound the sum of the strengths of the edges adjacent to vertex \( v \). (B) A worked example depicting edge weights in green and strength-index values in blue.
We use these definitions to describe a new property for vertex \( v \) in a network, the strength-index:

\[
s \equiv \max_{i \in \{d_v\}} \min \{s_i, r_i\}.
\]

Figure 2 illustrates the computation of \( s \). The \( i \)th edge incident to node \( v \) has the \( i \)th largest strength \( s_i \) and fractional rank \( r_i \). In this rectangle \( s_i > r_i \); so \( r_i \) is a candidate to be \( s \). We set \( s \) to the largest candidate over all such rectangles. By this definition, a vertex with strength index of 1.0 is in an isolated clique, while a vertex with strength index 0.0 participates in no triangles.

Let \( S \) be the sum of the strengths of all edges incident on a vertex \( v \) with strength index \( 0 \leq s \leq 1 \). Let \( A \) be the area under the curve in Figure 2. Because the \( x \)-axis has units of \( 1/d_v \), we have \( A = Sl/d_v \). The square in the figure has area \( s^2 < A = Sl/d_v \). Therefore we have a lower bound on \( S \): \( s^2 d_v < S \). Define the symmetric strength component (SSC) of \( S \) to be this lower bound \( \text{SSC} \equiv s^2 d_v \).

In social networks we have encountered, \( S \) (and therefore SSC) appears to be practically bounded by a relatively small constant such as 30. In the strength plots in Figure 5, some vertices have much higher \( S \) but they are a tiny fraction of the total vertices, as shown by the probability density functions in Figure 3. They are also likely part of the nonhuman portion of the network.

2.1 Assumption and network structure

We highlight the main assumption we use in our analysis. This is supported by the empirical evidence we just described. We then describe other network structure we enforce or that is necessary for our analysis to work well.

Main Assumption (A1). For a human vertex in a social network, the sum of the strengths of the edges adjacent to the vertex (\( S \)), is bounded by a constant.

For our method to be effective for partially cleaning a network, SSC should be a good approximation of \( S \), say within a constant factor, for most of the vertices. In publicly available social networks, the SSC empirically approximates \( S \) for most nodes, as we see via their respective probability density functions in Figure 3. We only consider edges that are reciprocated. We believe this represents more human social interactions. If we start with directed networks we preprocess them by removing nonreciprocated relationships.2 Note that this removes asymmetries introduced by follower or friend limits, or by the phenomenon of celebrity.

The concept of strength index is useful in two primary contexts. First, it is a way to quantify plausible human activity on the incident edges of a vertex. For example, a celebrity who uses some electronic social network may have millions of adjacencies. It is not humanly possible to put the necessary time and effort into cultivating and maintaining so many relationships. However, it is likely that some reasonable number of these relationships are true, meaningful ones. The strength index helps us find such relationships if they exist.

Secondly, we can model strength-index data with a simple formula. This model appears to fit the data of all social networks we have examined and is an especially good fit for the Twitter-2010 followers’ network. It also suggests a procedure for identifying non-human behavior. Let constant \( D \) (for “Dunbar”) be the bound on \( S \) in Assumption A1. The value of \( D \) may be network dependent. It depends on what the edges mean, what the natural human limitations are for that kind of action, and perhaps the properties or personalities of the people in that particular network. Therefore we compute a constant \( D_G \) for network \( G \). We omit the \( G \) when it’s clear from context.

Definition 1. The generalized Dunbar number \( D_G \) for social network \( G \) is a network-specific upper bound on the typical

\[2\]This makes the networks smaller, but that is not a motivation. Our methods scale to graphs with billions of edges. Including nonreciprocated edges does not change the nature of our results.
FIGURE 4 Twitter-2010’s strength-index scatter plot: (A) original graph and (B) after cleaning removes 5.6% of the reciprocated edges. Since many points co-occur, we color each point by its frequency, given by the heat-map index on the right. The green curve, from Inequality (2), marks the boundary between violators and nonviolators. The cleaning does not produce (B) merely by construction, as it would if we removed all violator vertices. The removal of violator-violator edges removes the implausibly dense subnetworks in (A), but leaves the violator-to-nonviolator edges that constitute roughly 30% of plot (A).

sum of incident edge strengths on any node in a human subnetwork.

As described before, the SSC is a lower bound on \( S \) the sum of the strengths of vertex \( v \)'s edges. We can generate a simple upper bound on the strength index given our assumption that \( D \) exists. We have \( D \geq S \geq s^2d_v \), and via simple algebra:

\[
\sqrt{\frac{D}{d_v}} \geq s \tag{2}
\]

This model gives an upper bound on the strength index of a human node as a function of its degree. Low-degree nodes are allowed to have a higher strength index. This is because having only a few edges, all strong, is plausible. Many people join Twitter, connect to a few close friends or family, and then stop using Twitter. Also, low-degree nodes with high strength index do not connect strongly to the rest of the graph and are therefore unlikely to have a large effect on graph-algorithm output.

See Figure 4 for the strength index plots of the Twitter followers’ network, before and after the cleaning process described below. The data points in these plots are (degree, strength index) pairs, colored by their frequency of occurrence. For example, for Twitter, at least 200 vertices of degree 50 have strength index 0.1. Also in Figure 4, note that Inequality 2 (the green line) seems to model a real phenomenon in the data. The violator vertices are those above the green line, and note that it is possible to be a violator with degree far smaller than the classical Dunbar number of 150.

Figure 5 shows edge-strength distributions and strength index data for 15 networks from SNAP [24], and Table 1 contains statistics for these networks. For our experiments we set \( D_G \) of network \( G \) to be a value close to the largest sum of edge strengths of any vertex in the network. We use the 99.85th percentile of the distribution in this paper. These values are included in Table 1.

We overlay the strength-vs-degree data for Twitter in Figure 4. We show the same plots for many other networks in Figure 5. The plot for each graph has a green curve depicting our model (2) of the strength-index upper bound for \( D_G \), for the given graph \( G \). Points located above the curve indicate nodes we call “strength-index violators” or simply “violators.”

To clean the network of nonhuman activity based on Dunbar’s limit [16], we use the above \( D_G \) value and remove the edges of the subgraph induced by the strength-index violators. In Section 4, we describe experiments validating this cleaning process. An obvious alternate process would be to remove the violator nodes themselves. However, doing this can remove a significant fraction of all edges in the network. Violators constitute only 0.04% of the nodes in Twitter-2010, yet removing them removes about 30% of the edges. Since we do not claim to classify individual nodes as human or non-human, we remove only edges of violator subnetworks.

In Figure 5, the networks in Row 1 are social networks where nodes largely share personal news or opinions. Networks on Row 2 involve people but perhaps with different activities. For example, the interactions involved in critiquing someone are different from friendship dynamics. Gowalla involves more action, so the dynamics might be different. In particular, this network can be used to arrange physical
FIGURE 5  Aggregate edge strength per-vertex distributions, and strength-index heat maps. See the caption of Figure 4 for an explanation of the heat maps. The upper figure for each graph shows the frequency histogram for aggregate edge strength on a vertex. Note the distinctive signature of a large social network in the strength-index plots (YouTube, LiveJournal, Pokec, Friendster, and Orkut), and the lack of clear distinction among most of the edge-strength distributions across all networks. The networks on the upper row are social networks. The networks in the middle row involve people, but are not typical social networks. See the discussion in Section 2. The networks on the bottom row are not social networks. See Section 6 for a discussion of some observations, conjectures, and puzzling aspects of these plots.

meetings. A Gowalla user may only list friends that live in the same city. The web-BerkStan network is a university web-link graph, but it is so old that it represents a time when users added links by hand, much as people friend in Facebook today. The last row has networks with no expected human activity. Note that there can be large differences in the strength values for the strength-index plots even when the edge-strength distributions look similar.

We will comment more on these visualizations, including conjectures about possible human “signatures,” in Section 6. For now, note that these plots immediately tell us that there are systemic violators in both Twitter and LiveJournal, yet these seem to be different phenomena. The Twitter violations are associated with high-degree vertices in a concave cloud, while the LiveJournal violations arise from structures on relatively low-degree vertices such as a nearly isolated 200-clique.

The Friendster network is unusually clean. The shape seems to fit the green curve well. Our data set is an archive of the Friendster network when it switched
from being a “normal” social network to a social gaming platform in 2011 [2,29]. It was primarily popular in Asia. It is unclear why Friendster did not develop the kind of unnaturally dense subgraphs that other social networks did.

### 3 | THE IMPACT OF CLEANING ON GRAPH PROPERTIES

In this section, we describe the of our cleaning process impact on two graph properties: PageRank and clustering coefficients. These differ enough in the cleaned graphs that we consider them to be different datasets.

Once we have selected a $D$ value, Inequality (2) partitions the data as shown in Figure 5. As described above, to clean the network, we remove all edges between vertices with strength index $s$ and degree $d$, such that $s > \sqrt{D/d}$.

#### 3.1 | Potential impact

In this section, we give evidence that if researchers are running a graph algorithm on a network they would like to be human, and they do not clean out the behavior we have described, it could significantly impact their results. Also, algorithm designers and implementors sometimes make decisions based upon structure they expect in their input instances. In particular there are algorithms designed to exploit social-network properties. The theoretical and/or empirical performance analysis will only be true to the extent that real networks match the assumed structure. Thus if nonhuman behavior affects important structural properties, it can affect algorithm/code performance.

We show that the nonhuman behavior we find has a significant impact on both PageRank and clustering coefficients.

#### 3.2 | PageRank

Figure 6 illustrates the decrease in PageRank for violators in the cleaned Twitter graph compared to the original. There were 16,701 violators, 819 of which were “superviolators,” nodes that had no remaining degree after cleaning. We include the latter with the maximum percentage change of any node that retained edges.

In the cleaned Twitter graph, 34% of violator vertices have a cleaned PageRank no more than half of the original value and 14% of them have a cleaned PageRank no more than a fifth of the original value. No appreciable percentage of the nonviolators experiences such a decrease. We conjecture that the creation of these violator subnetworks was a nonhuman process designed to artificially boost the PageRank importance of their members. In Section 4, we give further structural evidence for this claim.
3.3 Clustering coefficients

The global clustering coefficient (also called transitivity) of a network is the fraction of wedges in the graph that close into triangles (see Figure 1B). The clustering coefficient of a node is the fraction of that node’s wedges that close into triangles. A graph’s clustering coefficient is a measure of (non)existence of community structure. Kolda et al.’s [20] block two-level Erdős-Rényi (BTER) graph generator accepts as inputs a degree distribution and a per-degree clustering coefficient distribution. The latter is the average clustering coefficient for all vertices of a given degree. BTER produces graphs that empirically match both of the input distributions.

Figure 6 shows that clustering coefficients of the cleaned Twitter graphs differ considerably from the original. Thus giving these different distributions as input to the BTER generator will generate quite different random graphs.

4 INFORMALLY VALIDATING THE CLEANING PROCESS ON TWITTER

While the detection of bots in Twitter is not our goal, we now present two preliminary validation studies to increase our confidence that the edges removed by our cleaning process are most likely nonhuman. Recall that our cleaning is done based on topology alone, with a process inspired by evolutionary psychology. We do not appeal to Twitter account features such as tweets or account attributes typically used by machine learning classification approaches. This is because most researchers do not have access to such features in the openly available non-Twitter data sets. Yet researchers need cleaned versions of these networks for testing their human-network-only algorithms.

4.1 Ground truth comparisons

Our first validation step involves ground truth graciously provided by James Caverlee’s group at Texas A&M which comes from their work luring content polluters on Twitter [23] using honeypots (decoys). This ground truth consists of roughly 40,000 Twitter user ids classified as either “legitimate user” or “content polluter.” The content polluters self-selected by interacting with honeypots while TAM researchers sampled the legitimate users randomly. The Botometer team used this data set, which we will call TAM, to help validate their classifier [14,18]. Botometer can compute a score between 0 (totally legitimate) and 1 (totally bot) for any given Twitter user. In supervised learning experiments, Botometer achieved a classification accuracy of 95% (area under the Receiver Operator Characteristic [ROC] curve) for classifying legitimate users and content polluters in the TAM data set.

A key challenge with any validation of our Twitter-2010 cleaning is that the topology on which it relies is nearly a decade old. However, many of the nodes from that network are still active on Twitter at the time of this writing. Of the roughly 20,000 violators we find in Twitter-2010, roughly 8000 are still active. Since there is a fairly small overlap between our test sets and the TAM sets, we use Botometer to compare them. We used the Botometer API to compute scores for these violators, and for the TAM data set.

Figure 7 shows empirical cumulative distribution functions (CDF) for the Botometer scores of three sets of vertices: the TAM legitimate users, the TAM content polluters, and our strength-index violators. If our cleaning process had randomly selected legitimate accounts, we would expect a CDF more similar to the TAM legitimate users than to the polluters. However, the distribution nearly tracks that of the polluters.

The TAM content polluter ground truth was painstakingly obtained over 6 months by experts, and is likely to be nearly an ideal set with respect to the Botometer CDF. Our set of
Legitimate users and content polluters, and our strength-index violators (S-EK). More than 80% of both content polluters and strength-index violators have Botometer scores above 0.5.

We compared our sets of violators and nonviolators with the TAM content polluters and legitimate users. Since these were completely independent experiments in a graph with 40 million vertices, we did not expect a large overlap. However, there are 257 Twitter accounts present in both our dataset and TAM. This is the confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>Content polluters</th>
<th>Legitimate users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violators</td>
<td>223</td>
<td>33</td>
</tr>
<tr>
<td>Nonviolators</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

There are 33 nodes we consider violators, but TAM considers legitimate. We appealed to the average Botometer scores of these nodes to provide more context. This average is 0.64, only slightly less than the Botometer average score of 0.77 among the 223 nodes TAM labeled content polluters and we consider violators. Our strength-index analysis revealed subnetwork structures not reflected in any of the node-specific features used by the TAM and Botometer groups. In particular, it sometimes involves large quasi-cliques\(^3\) that would not be visible from studying nodes in isolation. Therefore it is possible that these 33 violators not tagged as content polluters nevertheless were adjacent to edges that do not represent legitimate human relationships. As one piece of anecdotal evidence, we looked at the Twitter accounts of a few nodes in such a quasi-clique. They appeared to be unrelated legitimate users such as companies in disparate businesses and celebrities. We conjecture this particular connected group of violators hired the same web management company which then connected them.

\(^3\)An \(\alpha\)-quasi-clique is a (sub)graph containing an \(\alpha\) fraction of the possible edges. We consider \(\alpha\) to be constant independent of the graph size.

There are only three nodes that TAM labels legitimate and we also label non-violators. This is a predictable artifact of sampling. While both sets selected violators/polluters intentionally, both selected nonviolators/legitimate users more randomly. We selected an arbitrary sample of roughly 10,000 of the 40 million Twitter-2010 vertices to query via the Twitter API, while TAM sampled roughly 19,000. We expect little overlap when the sample sizes are so tiny relative to the total number of vertices.

4.2 Follower list sequences

The distribution of violator Botometer scores shows that our cleaning process is removing connections largely associated with nodes that have Botometer scores above 0.5. This is encouraging, but we performed an additional validation step to gain more confidence.

Suppose an entity adds a fully connected network of content-polluting, automated nodes to Twitter. If the entity adds followers with an automated process, then they may well be added in the same order for each follower, by running a program with a fixed-order list. Each pair of adjacent nodes would agree nearly completely on the order in which they added their common neighbors. This could also happen if a web-management company had all its clients follow each other, connecting a new client to the clique in approximately the same order as the previous new client.

Let nodes \(u\) and \(v\) be adjacent violators in a dense subgraph. Let \(W\) be the set of common neighbors for nodes \(u\) and \(v\). Let \(N_u\) be the set \(W\) sorted by the time node \(u\) first started following each node in \(W\). Let \(N_v\) be the same sorted list for node \(v\). Then, if the edges in this dense subgraph were added in a consistent order, we would expect the common subsequence of \(N_u\) and \(N_v\) to have length \(\Theta(|W|)\).

Consider instead a community of friends in Twitter who join over time and follow one another in an order determined by convenience or when they meet. We would expect the order of adding followers in such a community to be closer to random.

If neighbors are added randomly, then it is unlikely there will be significant overlap in the order in which common neighbors are added. To see this, consider comparing the agreement between two random permutations of 1 \(\ldots\) n. Let one of the sequences be sorted without loss of generality. The longest common subsequence between the pairs becomes the longest increasing subsequence problem for a random permutation. Classical work shows that the expected longest increasing subsequence in a random permutation of length \(n\) approaches \(2\sqrt{n}\) as \(n\) grows large [3].

We used the multithreaded code of ref. [27] to efficiently find the largest clique among the violators and the two largest cliques among the non-violators in the Twitter 2010 data set. Using Twitter API, we retrieved the sequence of followers

\[ \text{FIGURE 7} \quad \text{Validation set comparing Botometer score} \]

distributions for Texas A&M legitimate users and content polluters, and our strength-index violators (S-EK). More than 80% of both content polluters and strength-index violators have Botometer scores above 0.5.

\[ \text{Cumulative Distribution Function} \]

\[ \text{Botomot Score \quad 0.0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1.0} \]

\[ \text{Nonviolators} \]

\[ \text{Violators} \]

\[ \text{TAM Legitimate Users} \]

\[ \text{TAM Content Polluters} \]

\[ \text{Topology ([S-EK])} \]
for the largest clique in the violator subgraph (318 nodes) and the second largest clique in the non-violator subgraph (53 nodes). The largest nonviolator clique was actually a system of nonhuman weather reporting accounts specific to cities that likely grew into a clique of violators in later years, so we discarded it. The API gives us only current information. In 2017 when we pulled these data, the largest remaining sub-clique of the violator clique had 164 nodes, and the largest remaining subclique of the nonviolator clique had 29 nodes.

We tested the hypothesis that a violator clique of size $n_v$ has pairwise longest common subsequences of followers of average length $\Theta(n_v)$, while the nonviolator clique of size $n_{nv}$ has such subsequences of average length $O(\sqrt{n_{nv}})$. Figure 8 shows the results. The violator clique does appear to have typical pairwise longest common subsequences of length $O(n_v)$ (highly unexpected of a human clique), while the distribution for the nonviolator clique has maximum density near the $2\sqrt{n_{nv}}$ expectation.

We acknowledge that the nonviolator result is not conclusive since at small $n$, $O(n)$ and $O(\sqrt{n})$ can be close together even with small hidden constants. However, we believe that the violator clique result is compelling, and if we had the Twitter-2018 follower network we could find a larger dense subgraph of nonviolators to increase confidence in the non-violator result.

4.3 Comparisons to other ways to clean networks

The validation results of Section 4.1 indicate that the sets of vertices our method selects as violators in Twitter-2010 have statistical properties similar to those of a set of content polluters selected by months of human effort. We now consider several other measures computable from topology only. If any reflect a social load placed on the brain of human vertices, these measures could be used in place of Equation (1) for computing vertex strength $S$ in our method.

We consider two types of alternative cleaning methods. The first type identifies violator vertices that are outliers with respect to the following basic attributes: degree, clustering coefficient, or PageRank. The second type uses the method we have described, but uses a different edge-strength measure $S$. We compute the total strength of all edges incident to vertex $v$ using the new strength measure, and then determine violators via inequality (2) as before.

The first alternative edge-strength measure is inspired by PageRank. Let $P(v)$ be the PageRank of vertex $v$. Define the unnormalized weight of edge $e = (u, v)$ as $P(u)/d_u + P(v)/d_v$, then normalize these weights to be in $[0, 1]$. We call this cleaning method S-PR for strength index with PageRank-inspired edge strengths.

The second alternative edge-strength measure is closely related to the Easley-Kleinberg measure in Equation (1). Berry et al. [4] defined the following measure of edge strength of edge $e = (u, v)$ for community detection:

$$s(e = (u, v)) = \frac{2t_e}{d_u + d_v - 2},$$

where $t_e$ is the number of triangles on edge $e$. This measures the fraction of edges adjacent to $u$ or $v$ that participate in a triangle with edge $e$. Their measure more generally considers an edge adjacent to either $u$ or $v$ “good” if it connects back to the other endpoint of $e$ via a “short” path, such as a path of length 4. But computing cycles longer than three in a social network is expensive. So they only used triangles in ref. [4]. Both measures (1) and (3) ignore the relationship between $u$ and $v$ because they look for other indications of a strong tie beyond the existence of the edge.

These two triangle-based measures are closely related. Let $b_e$ be the number of edges adjacent to either $u$ or $v$ that do not

FIGURE 8 Distribution of pairwise longest common subsequence lengths, along with the theoretical expectation, in (A) a 29-clique of nonviolators, and (B) a 164-clique of violators.
FIGURE 9 The top three panes show CDFs of three basic strategies for identifying violators: removing outliers with respect to (A) vertex degree, (B) PageRank, and (C) clustering coefficient. In (D) we use Dunbar intuition to constrain the set of vertices to those with degree at least 149, then remove clustering coefficient outliers. In (E) we use PageRank-inspired edge weights, then invoke Equation (2). The poor result suggests that PageRank-derived edge weights do not relate to Dunbar’s number as well as triadic closure methods do.

participate in a triangle on edge \( e \). Then the Easley-Kleinberg measure is

\[
\frac{t_e}{t_e + b_e - 2} \quad (4)
\]

and the Berry et al. measure is

\[
\frac{2t_e}{2t_e + b_e - 2} = \frac{t_e}{t_e + (1/2)b_e - 1} \quad (5)
\]

Both measures penalize an edge for having many adjacent edges that are not closely related to both endpoints by a triangle. These “distraction” edges pull the attention of vertices \( u \) and \( v \) away from each other. But the Easley-Kleinberg measure penalizes them more. We call the cleaning method that uses Equation (3) for strength index S-B.

Following Section 4.1, we use the TAM content polluter ground truth to evaluate the alternatives. We consider three different types of comparison, all based on the CDF of Botometer scores for the violator sets in question: visualizing the CDFs, evaluating their pairwise Kullback-Liebler divergence [22], and comparing false positives. We consider a false positive to be a vertex selected for removal that has a Botometer score of less than 0.5.

Figure 9 shows the CDFs of the Botometer scores of each violator set. We consider a method good if it closely matches the CDF of the TAM content polluters. In the top three sub-figures, we remove the outlier vertices (99.85th percentile) with respect to (a) vertex degree, (b) PageRank, and (c) clustering coefficient. In subfigure (d), we limit the candidates for removal to vertices with degree above 149, that is, only those with degree at least the traditional Dunbar number. Using this intuition dramatically improves the result, but still does not match the quality of our method. Subfigure (e) considers S-PR, computing strength index use the PageRank-inspired weights. We see a dramatic decrease in solution quality for S-PR.

Table 2 provides a different view of the same results. We compute the Kullback-Leibler (KL) divergence between the Botometer CDF of each violator set and that of the TAM content polluters. KL-divergence is an entropy calculation, and a score of 0.0 implies a perfect match of distributions. Our methods (S-EK and S-B, which are basically equivalent) dominate the others with respect to KL-divergence. We also count false positives (vertices selected for removal despite having a Botometer score of below 0.5) and find that the triangle-based methods most closely match the TAM content polluter’s rate.

Notably, the set of removal candidates that results from basing our process upon PageRank-inspired edge weights has a Botometer CDF that almost matches that of the TAM legitimate users. We use the latter as a set of “maximally bad” vertices to remove.

The triangle-based edge weights are connected to individual social effort, since a person is likely aware of local, two-hop connections between their own friends. This is consistent with the Dunbar-based motivation for our method. However, we believe that a PageRank style edge-weight cannot be tied to a vertex’s social effort since PageRank incorporates nonlocal information. In particular, the PageRank of an edge depends on the probability of traversing the edge in a stationary distribution, which could depend critically on topology far away from either endpoint of the edge.
TABLE 2  Kullback-Leibler divergences and false-positive rates (smaller is better). The TAM legitimate users are included as a maximally bad set of vertices to remove

<table>
<thead>
<tr>
<th>Method</th>
<th>KL-divergence w/TAM content polluters</th>
<th>Botometer false-positive rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAM content polluters</td>
<td>0.0</td>
<td>10.2</td>
</tr>
<tr>
<td>S-EK (s-index with Easley-Kleinberg weights)</td>
<td>0.1148</td>
<td>17.3</td>
</tr>
<tr>
<td>S-B (s-index with Berry et al. weights)</td>
<td>0.1154</td>
<td>17.3</td>
</tr>
<tr>
<td>S-PR (s-index with PageRank-inspired weights)</td>
<td>2.19</td>
<td>71.3</td>
</tr>
<tr>
<td>Clustering coefficient (Deg ≥149)</td>
<td>0.40</td>
<td>27.6</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.58</td>
<td>41.0</td>
</tr>
<tr>
<td>Degree</td>
<td>0.62</td>
<td>40.6</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.74</td>
<td>44.5</td>
</tr>
<tr>
<td>TAM legitimate users</td>
<td>2.47</td>
<td>91.6</td>
</tr>
</tbody>
</table>

5  | RELATED WORK

To the best of our knowledge, there are no other results on removing nonhuman elements from social network data sets using only topology. However, there is considerable work trying to detect bots in electronic social networks, including Botometer [14] and DeBot [10]. Bot detection systems try to find nodes behaving badly. We are trying to find nonhuman interactions that might not be malicious. In fact, in the case of the clique of Twitter nodes we found, the Twitter users may have been unwitting participants in the extra edges.

There is a recent study in the popular press similar to our follower sequence analysis of Section 4.2 [12]. In this remarkable article, the authors find structure in follower sequences beyond the simple subsequence analysis we present. They relate follower sequence rank to date followed and find clear patterns of automated activity. To act on this analysis, however, would require global Twitter API usage far beyond the current query rate restrictions.

There is other related work trying to find graph structure particular to social networks that enables more efficient algorithms. For example, Berry et al. [5] used the connected component structure of social networks, that is, the existence of a giant largest component and a logarithmic second-largest component. They compute s-t connectivity on a graph distributed among autonomous data centers while communicating provably fewer bits than the communication lower bound for general graphs. As Berry et al. argue, the giant component has a social explanation exists empirically and fits several theoretical random graph models. Other work posits properties at least somewhat (initially) supported empirically. The most famous is the power-law degree distribution [1,11,17]. Some researchers now believe the log-normal distribution is a better match for degree distributions [28].

Brach et al. [8] describe efficient algorithms for power-law degree networks that also follow a new Power-Law-Bounded-Neighborhood (PLBN) assumption. Roughly, the PLBN assumption is that for some parameter \( \alpha > 2 \), for every vertex \( v \) of degree \( k \), the number of neighbors of \( v \) of degree at least \( k \) is \( O(\log n + k^{1-\alpha}) \). Brach et al. show empirically that many networks follow the PLBN assumption.

5.1  | The strength index and the PLBN Assumption

Our primary assumption A1 is incomparable to the PLBN assumption in the sense that neither assumption implies the other. We now sketch this argument. To see that SSC does not imply PLBN, consider a tree rooted at a node \( v \) that has \( k \) children, and each of these children has \( k \) children. Note that this tree obeys the SSC assumption, but not the PLBN. Next, consider a network consisting of a node \( v \) with \( k \) neighbors, where these neighbors are partitioned into \( k/\log n \) subsets, each of which is connected in a clique. Note that this network follows the PLBN assumption. However, each edge incident to \( v \) in this network has weight \( (\log n)/k \), and so the total weight of edges on \( v \) is \( O(\log n) \). Hence, the network violates our assumption that the SSC is bounded by a constant.

6  | OBSERVATIONS FROM STRENGTH-INDEX VISUALIZATIONS

The strength-index plots of Figure 5 suggest two informal properties that distinguish social networks from others. In future work, exploring such plots might lead to a better understanding of networks.

First, social relationships (defined informally to be frequent highly clustered relationships among bounded-degree nodes) appear as red, vertical, oblong structures in the strength-index plots. This appears to be concentrated social behavior and is missing from nonsocial and nonhuman networks. For example, this feature is clear in Twitter's plot in
Figure 4, and in the other large social networks with friendship relationships. It is notably absent from human networks in which people cite or critique each other such as Epinions, reproduced in Figure 10, slashdot and, for more mysterious reasons to us, even Gowalla (which supposedly had friendship relationships).

Second, diversity, defined informally as a tendency to cover the entire space of humanly possible (degree, strength index) pairs in strength-index plots, appears in the large social networks (Twitter, LiveJournal, Friendster, Orkut, Youtube). As with social relationships, though, diversity is not present in the human, but nonsocial, networks represented in the middle pane of Figure 5 and in Figure 10. For example, consider vertices of degree 260. None of the possible strength index values are represented in Figure 10, while all nonviolating strength index possibilities are represented in the networks of the top (social network) pane of Figure 5.

Figure 5 leaves us with some puzzles:

- Pokec and LiveJournal appear to have full diversity until a strict degree cutoff (roughly 200 for Pokec and roughly 700 for LiveJournal). This is likely related to administrative degree limits. Notably, for example, LiveJournal had a limit of 750 friends in 2006 according to suggestions, livejournal.com. Perhaps Pokec and the other networks once had these cutoffs as well, though any follower limits in Twitter have not affected diversity.
- What is the mechanism for LiveJournal violators among relatively low-degree nodes?

7 | DISCUSSION

7.1 | The generalized Dunbar number

In this section we give a definition and justification for a generalization of the Dunbar number. While human capacity is constant, the amount of attention required for edge creation and/or maintenance in electronic social networks varies widely. This leads to our conjectured network-specific version of the Dunbar number.

Dunbar’s original bound on the size of a meaningful community of people is based on studies with chimpanzees. A wild chimp lives constantly within his/her social group. Dunbar posits that to maintain sufficient harmony, each chimp should know the relationship between every pair of chimps in the colony. For example, if two chimps are likely to fight, then others may work to prevent them from getting too close to each other. Dunbar thus considers a community a group of individuals who (approximately) know not only the other members of the community but also their pairwise relationships. He determined the largest chimp community size \( c \), and estimated the largest functional human communities by scaling \( c \) upward by the ratio of a human brain size to a chimp brain size. Thus he assumed memory was related to brain size (number of neurons).

We modify this concept by generalizing the binary counts of neighbor pairs, which Dunbar assumed all exist, into continuous sums of edge strengths. We proposed a name and informal definition for this, the generalized Dunbar number, in Definition 1 in Section 2.

Dunbar’s argument above implies that any group with much more than 150 nodes cannot be a functional community. Furthermore, any group with much fewer than 150 nodes cannot be proved nonfunctional by this argument. The notion of a generalized Dunbar number allows us to overcome both of these difficulties.

Following Section 2, we let \( D_G \) be the 99.85th percentile of the distribution of vertex total edge strength. Table 1 and Figure 5 show that these numbers are typically much smaller than 150, yet they lead to models (Equation (2)) that fit the data well.

We now relate Dunbar’s concept to groups of people larger than the standard Dunbar number. Therefore, even real communities are only quasi-cliques at best. A quasi-clique is a graph on \( n \) nodes that has \( \alpha \left( \frac{n}{2} \right) \) edges, an \( \alpha \) fraction of the possible edges. When we double the number of nodes in a group (community) from \( n \) to \( 2n \), the number of pairs increases by a factor of 4. A node \( v \) now has perhaps twice as many neighbors, so the average strength of a connection is halved compared to the community of size \( n \). Thus as community size grows, it is plausible that the total strength of the edges adjacent to a node stays approximately the same, bounded by \( D_G \), which does not depend on \( n \).

We now argue that the cognitive load for a node \( v \) to monitor pairs that do not involve \( v \) stays constant even though their number increased by four times. Given the converse of the triadic closure property, we expect that edges are about half as likely to be created by an average wedge in the community.
If the edge exists, it is likely to be half as strong as in the smaller community. Thus the average strength (required attention) of an edge in the 2n-node community is approximately one-fourth the strength in the n-node community. Thus a 300-node community might still be functional by Dunbar’s criteria.

While Dunbar’s original methods would accept any isolated community of size at most 150, we can remove edges from communities that are strictly smaller than 150 if $D_G$ is small.

### 7.2 Social-looking nonsocial networks in SNAP

We also ran this analysis on AS-Skitter [26], an autonomous systems network from SNAP, and it appears surprisingly social. However, we read from the home website (caida.org) that the data set we had used is deprecated. When we ran this analysis on the patent citation network, it also appeared very social. We believe this is because citations are added to patents by humans. When patent authors look for other patents to cite, they begin with the most recent patents. They find other relevant patents in the citations of what they find and, if they choose to call them out explicitly, include those citations. Thus the citations have a triadic closure property that grows in a way that human networks grow.

### 7.3 More details on the structure of human networks and human limits

This section provides a little more background on Dunbar’s work and the structure of social networks to support the modeling decisions we made in this paper. In particular, it gives social justification for triadic closure, a key motivation for our edge-strength measure.

Dunbar’s research [16] supports the existence of smoothly varying strength of human relationships. He describes the range of social bonds one human generally has. A person $x$ can “know,” by face and name, up to about 1500 other people. The relationships are hierarchical, with a small inner circle who would help person $x$ unconditionally, a next layer that is a little less dedicated to $x$, and so on out to people who are just acquaintances. Although not formally quantified, the strength of the tie from $x$ to each person in his/her social circle progressively drops at each level of the hierarchy. Each new level has about twice as many people as in all previous levels.

Easley and Kleinberg give three reasons for the triadic closure property. As we defined in Section 1, this is the tendency for a wedge $u \rightarrow v \leftarrow w$ to close into a triangle via edge $(u, w)$ when node $v$ has strong ties to vertices $u$ and $w$. The first reason is opportunity. Because node $v$ knows, and presumably frequently interacts with, nodes $u$ and $w$, (s)he has opportunities to introduce $u$ and $w$ to each other. The second reason is transitive trust. Nodes $u$ and $w$ both trust node $v$. Therefore, they are likely to pay more attention to each other when introduced by a mutual trusted friend, especially if introduced for a reason that benefits both. The third reason is social stress, which may apply strongly to some demographics. If node $v$ is spending time with node $u$, then (s)he is not spending time with node $w$ and vice versa. This causes stress on both relationships. It is frequently less stressful for node $v$ to do things with both $u$ and $w$ than to exclude one of them.

All three reasons for node $v$ to introduce nodes $u$ and $w$ decrease with decreased strength of ties, which supports our conjecture for the inverse property. If node $v$ has a strong relationship to node $u$, (s)he may mediate a special meeting with weak connection $w$ to help $u$ with a pressing problem, such as finding a job. But such connections will be special events. Kossinets and Watts [21] corroborate the importance of average strength with an experiment involving students. They wrote “The likelihood of triadic closure increases if the average tie strength between two strangers and their mutual acquaintances is high.”

We now give some reasons for our choices for the other cleaning methods we tried in Section 4.3. Although removing nodes that are outliers based on degree was not effective, it is a simple and natural heuristic method to consider for removing “non-human” nodes. It is generally accepted that human social networks have heavy-tailed degree distributions such as power law or log normal (e.g. [11]), though the exact nature of these heavy tails remains hotly debated [9]. Thus many nodes have few edges/observations and some nodes have connections to a large part of the graph. These ultra-high-degree nodes are frequently bots, including helpful weather bots and companies. But high-degree nodes can be celebrities who, although human, have many fans who have no real relationship. We hope our more nuanced edge removal process keeps the real human relationships of celebrities.

We showed in Section 4.3 that cleaning based on clustering-coefficient outliers is not effective, even though this method improves when limited to higher-degree nodes. Prior to this testing, one might believe nodes with outlier values of clustering coefficients could be candidates for non-human behavior. Human bonding/friendship means a human network has nontrivial communities: nodes can be broken into smaller communities of size at most 150, we can remove edges from communities that are strictly smaller than 150 if $D_G$ is small.

As described earlier in this section, Dunbar arrived at his number (150) by scaling up chimpanzee social capabilities by the relative size of a human brain. This seems dangerously heuristic. Nevertheless, Dunbar’s number has been proposed as an explanation for historical phenomena ranging from the splitting of medieval villages at size about 150 to the
organization of ancient Roman military units to the size of modern corporate groups [15]. This motivates using the Dunbar number as our heuristic cut off for communities where we do not expect quasi-cliques.

8 CONCLUDING REMARKS

We hope that other researchers or companies who have access to complete social-network data will rigorously validate our cleaning process. We believe the evidence we have presented is sufficiently strong to merit this investigation. That is, we have showed that nonhuman behaviors exists in these networks and can have considerable impact on algorithm output. Upon validation, our cleaning process could help companies such as Twitter clean out badly behaving accounts, perhaps in conjunction with methods that take advantage of extra features from the accounts.

A new branch of research in algorithm design is exploiting the properties of social networks to improve the efficiency of graph algorithms. These properties are still not well-understood. We contribute to this understanding through our definition and study of the strength index. Our study is based on accepted evolutionary psychology and leads to an intuitive cleaning process that removes nodes based on topological features that are implausible in human activity. We applied this process to several electronic social networks and found them to be polluted with large amounts of nonhuman activity. Our two validation studies on Twitter suggest most of the network elements our cleaning process removes are nonhuman.

There is considerable research in finding dense subgraphs (e.g., communities) in social networks. Since the kind of nonhuman behavior we find is dense, our cleaning process is likely to have an effect on the output of dense-structure-finding algorithms, perhaps better revealing the densest real human communities.

Nonhuman agents in electronic social networks can change their methods, to avoid detection via our cleaning algorithms. Depending upon their goal, they will likely have to work harder to benefit, while also avoiding our detection methods.

ACKNOWLEDGMENTS

This research was supported in part by the Division of Computer and Network Systems of the National Science Foundation Grant 131880 and by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. We are grateful to Aaron Kearns for his help with our social media education and to Philip Kegelmeyer and David Gleich for helpful discussions. We thank Clayton Davis from the Botometer team for helping us use the Botometer API and Kyumin Lee from Texas A&M for help with their ground-truth set. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

ORCID

Jonathan W. Berry
https://orcid.org/0000-0002-1768-685X

REFERENCES

14. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, Botornot: A system to evaluate social bots, Proceedings of the


