Behavior Research Methods manuscript No.
(will be inserted by the editor)

`wisdom_of_crowds`: A Python package for social-epistemological network profiling

Colin Klein[1] · **Marc Cheong**[2] · **Marinus Ferreira**[3] · **Emily Sullivan**[4] · **Mark Alfano**[3]

Received: date / Accepted: date

Abstract The epistemic position of an agent often depends on their position in a larger network of other agents who provide them with information. In general, agents are better off if they have diverse and independent sources. Sullivan et al. (2020) developed a method for quantitatively characterizing the epistemic position of individuals in a network that takes into account both diversity and independence. Sullivan et al. presented a proof-of-concept, closed-source implementation on a small graph derived from Twitter data. This paper reports on an open-source re-implementation of their algorithm in Python, optimized to be usable on much larger networks. In addition to the algorithm and package, we also show the utility of using our package to multiply profiling a much larger Twitter graph, showing a divergence between two epistemically important but distinct senses in which subgroups can be part of an 'echo chamber.'

Keywords social epistemology · Python · testimony

1 Introduction

Most of what we know we know because we learned about it from other people. *Social epistemology* is the subfield of philosophy that studies how knowledge and justification depend on the testimony of others (Goldman, 1999). In recent years, social epistemologists have moved away from considering dyadic relationships between individuals to consider the ways in which social epistemic *networks* shape

Work on this paper was supported by ARC Grant DP190101507 (to C.K. and M.A.) and by Templeton Grant 61378 (to M.A.).

✉ Colin Klein
E-mail: colin.klein@anu.edu.au

[1] The Australian National University
[2] University of Melbourne
[3] Macquarie University
[4] Eindhoven University

the information we receive (O'Connor and Weatherall, 2019; Alfano and Sullivan, 2020). A focus on networks has been influential because it allows philosophers to connect their concerns to the substantial body of empirical and simulation work on real-world networks and their graph-theoretic properties.

Sullivan et al. (2020) presented a method for quantitatively characterizing the epistemic position of individuals in a network. Broadly speaking, individuals are in a better epistemic position if they are receiving information from *diverse* and *independent* sources, with the more diversity and independence the better.

Sullivan et al. operationalized these two concepts in a way that allowed them to provide an interesting profile of a small 185-member Twitter community. That work relied on a bespoke, closed-source codebase. As it was built as a proof of concept, it was also not optimized in ways that naturally scaled to larger networks. TThis made it difficult to apply the technique to other datasets, such as networks from other social media sites or networks created from artificial social simulation algorithms (e.g. Laputa (Olsson, 2011)).

To make this tool more widely available to researchers, we therefore present `wisdom_of_crowds`, a complete ground-up reimplementation in Python of the core Sullivan et al. (2020) concepts. The code is optimised to deal with larger networks. It also includes some standardized helper functions to allow for coordinating results between research groups. We have made the code for this package open source, under the Creative Commons *Attribution-NonCommercial-ShareAlike 4.0 International*: CC BY-NC-SA 4.0 license,[1] and available on GitHub. As much as possibly we have relied on open-source Python packages such as `networkx` (Schult, 2008) and `matplotlib` (Hunter, 2007) as they have been rigorously tested and are freely available for auditing and peer review. For reproducibility and good practice, the `pytest` (Krekel et al., 2004) library is used to provide a unit-testing framework.

This paper presents the core concepts and the details of the `wisdom_of_crowds` package, and describes a few sample results on both simulated and real-life large networks.

2 Methods

2.1 Core concepts

Consider an epistemic network G where nodes are epistemic agents and edges represent the relationship of receiving information via testimony. 'Testimony' is used broadly in social epistemology for any way in which one source delivers information to another, and includes speech, writing, and other forms of media.All things being equal, a node is better off receiving information from more and more diverse nodes. However, testimony is often transmitted in chains, and this transmission need carry only the content of the information, not information about the original source or the intermediate links. This complicates the position of any individual who is trying to learn from multiple sources. For example, piece of gossip heard from two people seems more reliable than from one, but that reliability is undermined if both heard it from the same person (Alfano and Robinson, 2017).

[1] Full license terms can be found at `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

Following Sullivan et al. (2020), say that a node n is an m, k-observer just in case it receives information from a set of at least k different nodes which are pairwise at least m steps away from one another, when considered on the subgraph of G that does not contain n. If G is directed, then candidate sources must be at least m steps away in both directions. The removal of n from consideration in the case of distances is necessary for directed graphs, as otherwise all sources to n are trivially at most 2 steps apart; we carry over that requirement to undirected graphs as well.

In this work, we assume $1 \leq m \leq 5$, as most real life networks have length 6 paths between most arbitrarily chosen nodes (Milgram, 1967). We bound $2 \leq k \leq 5$, because a node with a single source is in a very poor epistemic position with respect to diversity of input. Note that it is a consequence of the definition that if n is an m, k-observer, it is also both an $m-1, k$-observer and an $m, k-1$-observer (assuming $m-1$ and $k-1$, respectively, are permissible values).

Given this definition, the core concepts in Sullivan et al. (2020) are defined as follows.

$S(n)$ gives a measure of the independence of sources to node n. Consider the set s of possible pairs (m, k) for which n is an m, k-observer. Then define

$$S(n) = \begin{cases} 0 & \text{if } s = \emptyset \\ \max\{mk : (m,k) \in s\} & \text{otherwise} \end{cases} \tag{1}$$

In other words, $S(n)$ is just the largest mk such that the n is an m, k-observer. If S has 0 or 1 nodes as sources, they are considered as being in an epistemically bad position, and so $S(n) == 0$. Note that given this definition, possible S values do not increase smoothly. Given the bounds set out above, $S(n) \in \{0, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 20, 25\}$.

$D(n)$ measures the diversity of the sources that contribute information to n. Let each node i be associated with a set a_i of epistemically relevant attributes. These might be group affiliations, topics of interest, scientific approaches, political leanings, and so on. Let s be the set of n's sources. Then define

$$D(n) = \mid \bigcup \{a_i : i \in s\} \mid \tag{2}$$

That is, D gives the number of distinct *types* of information that feed into n.

Finally, as the epistemic position of a node is a function of both the diversity and independence of sources, we define $\pi(n) = S(n)D(n)$.

There are a few notes to make about the implications of these core concepts in real social networks. Regarding m, k-observers, while higher rankings are better and all the nodes with a specific value of m and k are members of an equivalence class, the framework does not posit which of two m, k-observers is better positioned if one has a higher m value and the other a higher k value—for instance, whether $2, 3$-observers are better or worse placed than $3, 2$-observers. The framework thus does not provide a total order but instead provides a collection of partial orders.

2.2 The package

The core of the `wisdom_of_crowds` is a class `Crowd`. `Crowd` is initialized with a *NetworkX* graph (encapsulating the social network's edges and nodes), and provides

various functions to calculate the metrics defined above. Much of the heavy lifting is done by the `Crowd.is_mk_observer(n,m,k)` function, which returns `True` just in case node n is an m, k-observer.

Calculating whether a node is an m, k-observer combines multiple shortest-path problems with clique-finding problems. Naïve approches to the latter have complexity $O(n^k)$ (Vassilevska, 2009). Given that we are considering unweighted paths, the shortest-path problem has a reasonably efficient linear-time solution, but the requirement to remove n from the distance calculations also means that network shortest paths cannot simply be calculated at the outset. In the worst case scenario, they must be recalculated for each pair of sources for each node.

Hence this is a computationally difficult problem to brute-force. In practice, efficient caching and testing of seen paths plus greedy k-clique algorithms means that worst-case performance can often be avoided for realistic networks. Paths are stored as part of the `Crowd` object, which often gives substantial speedups when batch processing over every node. A *pickle*d object will include a cache of paths and S values unless cleared beforehand. Recalculation of `Crowd.is_mk_observer()` also requires recalculating cliques, however, and for densely connected nodes that can take nontrivial additional time. As the envisioned use case for large graphs is for one-shot batch processing, we do not anticipate this being an issue. Nonetheless, the open-source nature of our code allows for trivial modifications (e.g. using the *multiprocessing* package) allowing it to work on distributed multiprocessing systems for substantial performance gains.

The m, k-observer functionality is the basis for calculating D, S, and π. D is calculated on node attributes, and users can specify the appropriate key for the attribute. If a single attribute is supplied, D is calculated using the singleton set containing that attribute.

In addition to the standard measures, we also define the `h_measure()` of a node n as the smallest x such that n is an x, x-observer (compare to the standard definition of the h-index in citation practices). Sullivan et al. (2020) suggest that being a 3,3-observer is the minimal secure epistemic position, and the use of a single non-multiplicative standard may be useful for some cases.

Finally, the package includes two helper functions to allow for comparable reporting and display across different users. `iteratively_prune_graph()` takes a *NetworkX* graph, removes small-degree nodes, small-weight edges, and takes the largest connected component in what remains, iterating this process until the graph is stable. The thresholds are parameterized; the default is for $indegree + outdegree \leq 1$ and no edge culling, as per (Sullivan et al., 2020). `make_sullivanplot()` makes a summary figure of a whole network in the style of Sullivan et al. (2020, Figure 7). It can produce standalone plots or return a subplot in a specified `matplotlib` axis.

3 Results

Figure 1 shows the efficiency of batch calculating S for each node of a `Crowd` on random graphs with varying parameters for probability of edge connection. (See appendix A for details).

As the log-linear plot figure 1 shows, there is a roughly exponential relationship between the number of nodes and runtime, with the exponent a function of the

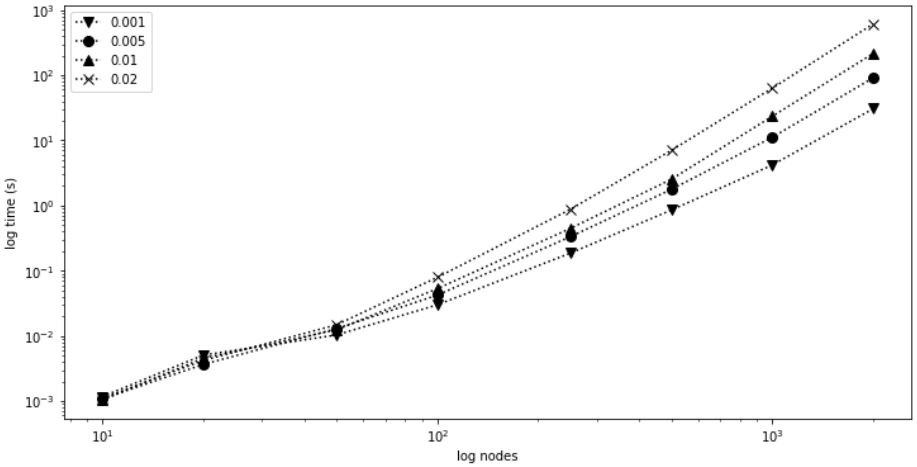

Fig. 1 Timing curve for random graphs by number of nodes. Different markers represent different connection probabilities for nodes.

number of edges. This suggests that the efficiency of our code approaches what would be expected given the fundamental complexity of the clique-finding problem. Note that the exponential growth means that the boundary between computationally tractable and intractable graphs can be relatively tight. Judicious pruning often makes a difference.

Figure 2 plots S, D, and π for a real-life network of participants who retweeted content around the Black Lives Matter movement in the first half of 2020 (see appendix A for details). Sullivan et al. (2020) used an earlier version of this dataset and were able to examine a network of 185 nodes. This analysis was run on a culled network of ∼16k nodes and ∼145k edges. Batch processing took about 6.25 hours on a 2017 desktop iMac.

We examine both the network as a whole and three identified subclusters in the graph. The left half of 2 shows S, D, and π for the network, where D is calculated via the subcluster identity of sources. The right half of figure 2 recalculates D and π based on a 9-topic NMF topic model of aggregated tweets (compare with the 3-topic model of Sullivan et al. (2020)).

Figure 2 shows the utility of profiling networks using our toolkit. On the left, one can see that Republicans appear to be in the worst epistemic position in terms of the other subgroups with which they interact: they have a generally low D, suggesting that they tend to listen mostly to in-group members. However, they have a relatively high S and therefore a π comparable to other subgroups. Compare this to the topic-wise graph, in which Republicans have a a relatively high diversity for *topics*, one at least as good as other groups. The activist group shows something of the inverse pattern. That is, they show a more varied range of S and D values for group-group interactions, but a comparatively lighter graph with fewer topics for the broad span.

These results might suggest that both groups are part of 'echo chambers' (Alfano et al., 2018; Nguyen, 2020), but in different ways: the right tends to be a

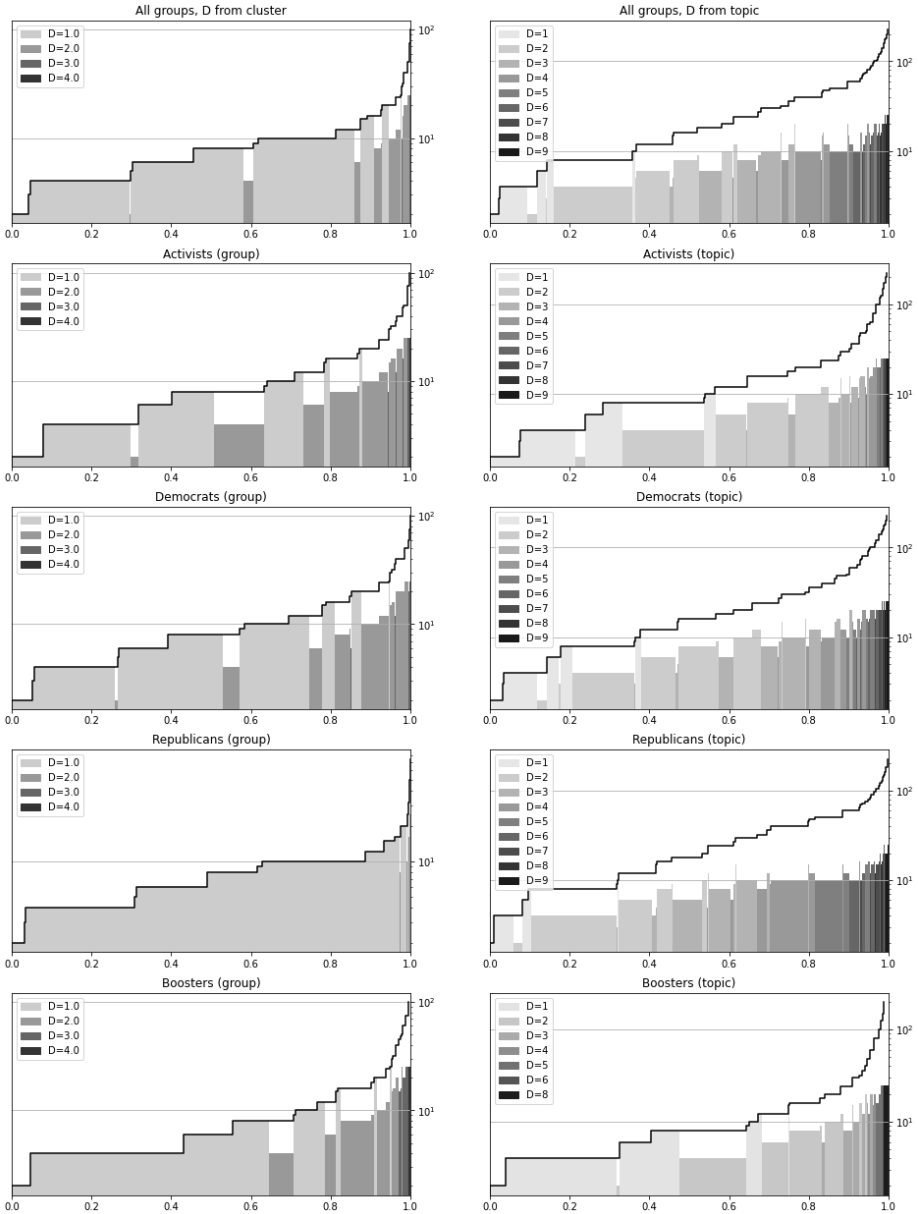

Fig. 2 Profile plots for entire network and subgroups looking at clusters (left) and topics (right). X axis is proportion of total, Y axis shows both S (height of bars) and π (black line), plotted on a log scale.

monoculture socially but a polyculture topically, with a converse pattern on the left.

Finally, we note that all subgroups, in both domains of measurement, tend to have an $S < 10$ for more than half the population. This replicates the observations of Sullivan et al. (2020)), in which most participants end up in a comparatively poor epistemic position. However, most groups tend to contain at least a small sub-population which is well-connected and often with a relatively high D. We note that this is especially the case with our 'Booster' group, a small subset who seemed to be mainly concerned with amplifying and relaying the content of other groups.

4 Discussion

Our results show that it is possible to replicate the methodology used by Sullivan et al. (2020) in larger networks, and that insights about the relative epistemic positions of different communities within a network can be drawn from plotting these parameters. As our package and its dependencies are all open source, this makes it possible for researchers in a range of fields (including philosophy, psychology, sociology, anthropology, communications, and network science) both to conduct new research and to reanalyze networks that they have previously studied.

So far, the only networks that have been studied using this tool are from Twitter (and, as part of our testing framework, *de rigueur* standard social networks such as the Florentine Families network of marriages (Breiger and Pattison, 1986)). We anticipate that future research will expand the types of social networks under study. Other sources from social media such as Facebook, Reddit, and YouTube all seem to be viable candidates for study. Considering offline epistemic networks would be especially valuable, as their structure may be interestingly different from the structures found online; as well as epistemic network simulations, created with tools such as Laputa (Olsson, 2011). We expect that studies of friend networks, organizational networks in industry and the military, networks of sources used by journalists, criminal cartel networks, and academic citation networks would prove valuable.

Moving beyond that, it would be interesting to study networks with more than one type of testimonial edge (e.g., public communications versus private ones). One intriguing hypothesis is that these may differ in structure even if they contain the same nodes, and that individuals who are central in public networks but peripheral in private networks (or vice versa) would tend to play unique roles in the social epistemology of those networks. For instance, someone who is privately in communication with a very large number of others but not publicly visible is in a position to exert influence because the others may assume that they have a much better epistemic position than they actually do.

The exploratory profiling made possible by our tool reveals patterns of epistemic isolation and interaction across real-world networks, and suggests possibilities for more specific analyses. By providing it to the community at large, we hope to facilitate further modeling of epistemic networks across a variety of domains.

A Supplemental methods for figures

Random directed graphs were generated using the `networkx` generator `fast_gnp_random_graph()` with the parameters indicated in the figure. Timing was done using the python `timeit` package over a single iteration.

The data collection for figure 2 was done as part of a project examining Black Lives Matter discourse on twitter. We queried the Twitter Streaming API with a series of Black Lives Matter (BLM)-related keywords, hashtags, and short expressions in a window between January and July 2020. We used a sliding window to take into account that between 80%-90% of retweets occur within 5-7 days, with diminishing returns beyond (Kwak et al., 2010). The dataset comprised ∼4.6M original tweets between January 13th and July 18th and ∼94.5M retweets from January 18th to July 23rd; these tweets were produced by ∼2.0M distinct authors. After the murder of George Floyd (May 25th 2020), the number of daily tweets increased by several orders of magnitude (from ∼255k to ∼4.35M).

We generated a *retweet network* (Sullivan et al., 2020), a weighted directed network where nodes are authors and the weight of an edge from node u to node v represents the number of times that user v retweeted user u. Self-retweets were discarded. Note that given this definition, users who retweeted but who did not author any tweets could not be nodes in the network. We took the largest connected component of this graph as the starting point for analysis. To find clusters, we used *igraph* (Csardi and Nepusz, 2005) and the Python *leidenalg* package which implements the Leiden community detection algorithm (Traag et al., 2019). We found first-level clusters using Modularity Vertex Partitioning, preserving clusters with more than 10% of the original nodes. This gave 4 clusters, covering 83% of the graph. Next, we manually inspected the 100 most-influential nodes within each group, characterizing the communities as Activists, Center-Left Democrats, Republicans, and a set of "Boosters" who mainly amplified the content of the first two groups.

Topic models were fit using *scikit-learn*'s non-negative matrix factorization (NMF). Documents were aggregated tweets of each author in the original graph, preserving those with an aggregate length of ≥ 50 words. Tweets were preprocessed to remove non-alphabetic content and common English stopwords. NMF was fit on a tf-idf representation of this corpus with `min_df=0.05`. We chose 9 topics because it was large enough to pull out an interesting variety of themes. The top five words of each topic are as follows:

```
0: matter life black live say
1: blacklivesmatter georgefloyd justiceforgeorgefloyd protest today
2: police brutality protest protester racism
3: blm antifa protest democrat trump
4: people white just say fuck
5: backtheblue alllivesmatter bluelivesmatter officer trump
6: george floyd breonna taylor justiceforgeorgefloyd
7: petition spread sign tweet donate
8: black live woman trans community
```

For further analysis, we used `iteratively_prune_graph()` with a node and weight threshold of 3, which resulted in a tractable subgraph with 16249 nodes and 145246 edges. This subgraph had very little representation from the 'booster' group, so they were omitted from further analysis. For the left side of 2, D was calculated using the cluster identity of the node. For the right side, D was calculated using the `argmax` of the fitted and normalized W matrix for the topic model. This gives the topic that is most distinctive of each user's tweets.

References

Alfano, M., Carter, J. A., and Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association*, 4(3):298–322.

Alfano, M. and Robinson, B. (2017). Gossip as a burdened virtue. *Ethical Theory and Moral Practice*, 20(3):473–487.

Alfano, M. and Sullivan, E. (2020). Humility in social networks. In *The Routledge Handbook of Philosophy of Humility*, pages 484–494. Routledge.

Breiger, R. L. and Pattison, P. E. (1986). Cumulated social roles: The duality of persons and their algebras. *Social networks*, 8(3):215–256.

Csardi, G. and Nepusz, T. (2005). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Krekel, H., Oliveira, B., Pfannschmidt, R., Bruynooghe, F., Laugher, B., and Bruhin, F. (2004). pytest 6.2.5.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161.

O'Connor, C. and Weatherall, J. O. (2019). *The misinformation age*. Yale University Press.

Olsson, E. J. (2011). A simulation approach to veritistic social epistemology. *Episteme*, 8(2):127–143.

Schult, D. A. (2008). Exploring network structure, dynamics, and function using networkx. In *In Proceedings of the 7th Python in Science Conference (SciPy)*. Citeseer.

Sullivan, E., Sondag, M., Rutter, I., Meulemans, W., Cunningham, S., Speckmann, B., and Alfano, M. (2020). Vulnerability in social epistemic networks. *International Journal of Philosophical Studies*, 28(5):731–753.

Traag, V., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233.

Vassilevska, V. (2009). Efficient algorithms for clique problems. *Information Processing Letters*, 109(4):254–257.