Ram and Rizoiu *EPJ Data Science* (2024) 13:53
https://doi.org/10.1140/epjds/s13688-024-00492-z

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

# Empirically measuring online social influence

Rohit Ram[1,2*] and Marian-Andrei Rizoiu[1]

*Correspondence:
rohitram96@gmail.com;
rohit@thaum.io
[1]University of Technology Sydney, Sydney, Australia
[2]Thaum, Canberra, Australia

**Abstract**

Social influence pervades our everyday lives and lays the foundation for complex social phenomena, such as the spread of misinformation and the polarization of communities. A disconnect appears between psychology approaches, generally performed and tested in controlled lab experiments, and quantitative methods, which are usually data-driven and rely on network and event analysis. The former are slow, expensive to deploy, and typically do not generalize well to topical issues; the latter often oversimplify the complexities of social influence and ignore psychosocial literature. This work bridges this gap by introducing a human-in-the-loop active learning method that empirically quantifies social influence by crowdsourcing pairwise influence comparisons. We develop simulation and fitting tools, allowing us to estimate the required budget based on the design features and the worker's decision accuracy. We perform a series of pilot studies to quantify the impact of design features on worker accuracy. We deploy our method to estimate the influence ranking of 500 X/Twitter users. We validate our measure by showing that the obtained empirical influence is tightly linked with agency and communion, the Big Two of social cognition, with agency being the most important dimension for influence formation.

**Keywords:** Active learning; Psychometrics; Online social networks; Social influence

## 1 Introduction

Social influence is a pervasive force that shapes our interactions with others and contributes to the emergence of complex societal behaviors. Understanding social influence mechanisms is crucial for comprehending how individuals and groups decide and act in concert. Empirical measurement of online influence can help us identify the factors that shape online behavior, such as the influence of opinion leaders, the latent topology of social networks, and the role of algorithms in directing the flow of information. This, in turn, assists us in developing strategies for managing and regulating online behavior and ultimately contributes to creating healthier and more sustainable communities.

Social influence is defined as *a change in a person's cognition, attitude, or behavior, which has its origin in another person or group* [1]. Influential people are those capable of effecting this change and here we measure this influence in people. This element of behavioral dynamics makes influence surprisingly challenging to quantify. There are two main approaches to achieving this, each with merits and shortcomings; psychosocial and quantitative. Psychosocial experiments are conducted in controlled laboratory environments [2–4], capable of investigating the minute nuances of the influence phenomenon but are consequently limited in size and cannot assess emergent behavior. On the other hand, quantitative methods for measuring influence based on online social media data [5] are often ad hoc, somewhat arbitrary, and do not relate strongly to social phenomena [6]. The insights from one domain remain largely orthogonal to the other [7]. Furthermore, the capability to apply psychometrics at an online scale is lacking. The question is, therefore, *how can the influence phenomenon be measured at scale, while maintaining a high-quality experimental environment?*

This paper explores how to measure social influence empirically and at scale by ranking a set of online social media users (hereafter called *targets*). We base our quantification on the perception of social influence by peers—we ask people to compare two targets and choose the more influential one. Note that, the peers are Amazon Mechanical Turk (MTurk) workers and targets are online users, who represent disjoint sets. We assume that people can judge the relative social influence between individuals. While this may seem like a strong assumption, sociometric research often utilizes peer-perceived measures of social attributes—such as popularity and reputation—and finds these measures are robust and reliable [8–10]; however, the methodology is rarely applied to the scale of our cohort. Influence measurement has two main challenges: scaling up the experiments to hundreds or thousands of targets and constructing the ranking from pairwise comparisons. The naive approach would require performing all possible comparisons – which scales quadratically with the number of targets. This quickly becomes prohibitive in a crowdsourcing environment where one pays a fixed price per comparison.

We solve both challenges by proposing a system with three components. First, we address the scaling challenge. As the *first component*, we use a human-in-the-loop active learning technique—human crowdsource workers are allocated tasks dynamically by an algorithm based on the previous decisions of other workers. Figure 1 illustrates our human-in-the-loop system. The active learning mechanism generates comparison queries and adds them to a queue. Multiple crowdsource workers concurrently complete batches

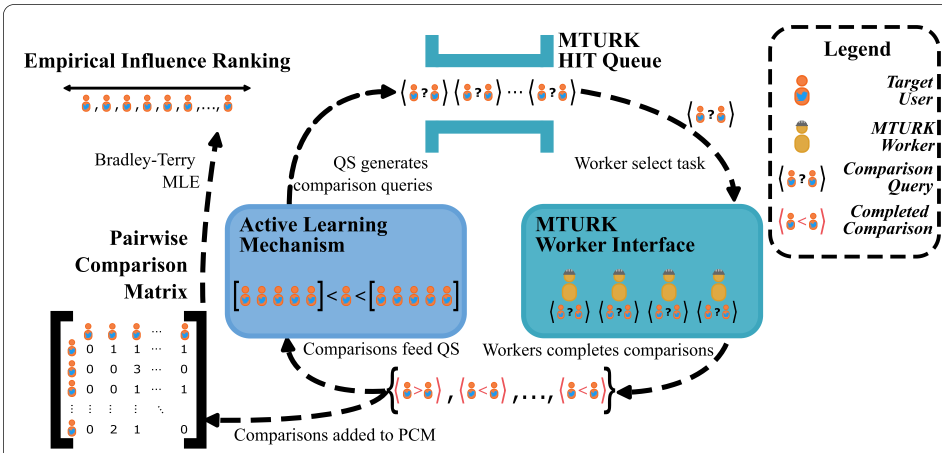


**Figure 1** The schema illustrates the human-in-the-loop system for generating comparisons

of queries—each query compares two target users and asks who is the more influential. The completed pairs are returned to the active learning mechanism, which uses them to generate more comparison queries, thus closing the loop. Section 3.1 further details our empirical influence measurement approach and shows it reduces the required pairwise comparisons and improves each decision's utility. For example, for 300 targets the naive approach would cost ∼$2500, but our approach would only cost ∼$67. Our framework scales loglinearly and, consequently, gains improve with scale. For 600 targets the naive approach would cost ∼$10,100, but our approach would only cost ∼$145.

Next, we address the ranking challenge. The *second component* is an augmented Bradley-Terry model [11] that we leverage to build an influence ranking from pairwise comparisons while accounting for systematic noise in worker decisions. We show that the systematic noise is theoretically linked to the expected worker decision accuracy, allowing us to recover, via simulation, the relationship between the required budget, the systematic noise, and the quality of the influence ranking. Here, *budget* refers to the number of pairwise comparisons we require MTurk workers to complete, which is directly proportional to monetary costs on the MTurk platform. This allows answering questions such as "What is the expected ranking quality for a given number of targets and my maximum budget?" Conversely, it can answer the question, "What budget do I require to achieve a certain ranking quality for my given number of targets?" With systematic noise approximated through experiments with real workers, we infer the ranking quality achievable.

The *third component* is an MTurk survey instrument that we develop and optimize to estimate the impact of design features on worker decision accuracy and reduce the required comparison budget. We account for several design features—such as user metrics, proxies, and qualifications—and show through a series of ablative pilot studies that improving the MTurk worker interface improves the accuracy of worker decisions. We find that each feature individually improves worker accuracy, and jointly using all features provides the best results.

We validate our empirical influence measure by linking it to the Big Two of social cognition (i.e., agency and communion). The Big Two are known as the fundamental dimensions of social comparison, and they are strong predictors of social factors, such as gender, class, and power [12, 13]. Furthermore, prior works [14–16] have linked the Big Two and social influence. It follows that this link should also be measurable between the Big Two and our proposed empirical influence measure. We build regression models for the empirical influence using the intensity of the Big Two as features. We find that both agency and communion are significantly linked to empirical influence. Agency (the drive *to get ahead*) appears more important than communion (the drive *to get along*) for influence formation in our X/Twitter users, given the transient and non-ongoing nature of relationships on the platform. Interestingly, the user follower count—a widely used proxy for social influence—does not have any detectable connection to social cognition.

*The main contributions of this work include:*

- A *crowdsource empirical influence measurement framework* to build the influence ranking of large cohorts that leverages human-in-the-loop active learning.
- A set of *simulation and fitting tools* for the measurement framework letting practitioners assess the required budget, design quality of annotation environment, and annotation fidelity.

- A *pilot study of MTurk design features* that improves the expected worker decision accuracy and minimizes the required budget.
- A showcase of the *link between empirical influence and the Big Two of social cognition*.

*Challenges of Measuring Social Influence.* Social influence is a complex behavioral phenomenon, which is inherently subjective. Our measure is based on influence as perceived by MTurk workers; while we enforce interventions to manage worker subjectivity, we do not remove the inherent societal biases around influence. For example, would a profound impact on a small group or a minor impact dispersed over a vast group display more influence? Furthermore, influence is often described with a social context and a goal. One might choose to measure influence by whether an influencee performs an influenced action (i.e., compliance). In this work, we limit our scope to the facet of social influence related to online opinion formation.

We assume that people have a latent influence factor, which describes their influence with respect to the population. Additionally, we assume that a total ordering of the social influence of online users exists. In our protocol, we do not necessarily compare every pair of target users, however, we infer a relation between every pair using this latent influence factor. Note, that our monadic interpretation of influence has differing applicability compared to a dyadic one. This interpretation allows for a macro-scale analysis of the attributes associated with influence in general. However, it does not allow for analysis of micro-scale interactions and the attributes of relationships that reinforce influence (e.g., homophily-induced influence).

## 2 Preliminaries and related work

We structure the discussion of related work and preliminaries into two parts. Section 2.1 presents previous influence estimation approaches, accenting crowdsourced quantification and psychometrics. Section 2.2 introduces the required concepts of pairwise comparisons, the Bradley-Terry model and how to construct a ranking.

### 2.1 Influence estimation and crowdsourcing

*Empirical Influence Measurements.* There are two popular avenues in literature to investigate the dynamics of influence. The first avenue within the sociology literature uses agent-based simulations, which can demonstrate sufficient conditions for particular emergent outcomes [6, 17–19]. However, this avenue has rarely correlated outcomes with empirical data, and is predominantly validated at a phenomenological level [18, 20]. The second (and more practiced) avenue emanates from psychology and tests for particular behaviors in a social context, such as social distancing [3]. Researchers aim to identify the characteristics that allow individuals to exert influence—such as authority, likeability, attractiveness, and expertise [21]. Neither avenue deals with measuring the relative influence between individuals. Our work uses human workers to perform pairwise comparisons and creates a ranking of a set of targets based on their human-perceived influence. The approach is inspired by sociometry—used to measure social quantities, such as status, popularity, and reputation [22]. Generally, sociometric methods elicit (positive and negative) nominations about the quantity of interest (e.g., status) from members within a group and derive a measure from these nominations. Our approach differs from traditional sociometry in its application in the online space and, consequently, to peers in a large population who do not necessarily know each other.

*Crowdsourcing Psychometrics.* Extant literature has used crowdsourcing for psychometrics to measure context relevance [23], factors of explainability [24], and interest [25]. Unlike our work, these studies treat workers as the subject and fail to scale using modern active learning methods.

*Influence Co-variates.* Influence has been related to several psychosocial attributes, such as likeability and authority [21, 26]. Given its central role in mediating human interaction, influence likely has an abundance of co-variates. Furthermore, these co-variates may be evident in macro-level processes related to influence, such as the formation of social movements. Two of particular relevance are *resource mobilization* capability and *engagement styles*. Bhattacharya et al. [27] show that resource mobilization is required for social movements to take-off. They utilise network analysis to show that communities coalesce around the same resources, and become more cohesive, in the lead up to a social movement. Furthermore, Shaik et al. [28] show that engagement with different media modalities (e.g., images or videos) varies throughout a social movement. They show that some engagement styles dominate in the initialization and amplification of social movements. While *resource mobilization* capability and *engagement styles* are likely related to social influence, there are challenges in operationalising them in setups such as ours. In both cases, characterisation of these co-variates are emergent in social movements (i.e., macro-scales), and it is unclear how to operationalise these for each user (i.e., a micro-scale), as in our setup. For resource mobilization, the construction of the user mention network, and subsequent network analysis, is infeasible at the scale of our dataset. Furthermore, there is generally insufficient activity at a user level and it is unclear how to classify the engagement styles of users on Twitter/X. Nonetheless, interaction of these co-variates with influence warrant investigation by future work. Given that we measure influence, as it is perceived by peers, we choose to investigate the big two of social cognition instead. Agency and communion have recently shown significant utility in explaining social phenomena.

*Crowdsourcing Influence.* The investigation of influence through crowdsourcing is not unique to our work. There are two main themes in the literature: recruitment to artificial social platforms, and identification of micro-influencers. In the first theme, the extant literature constructs artificial social platforms and recruits crowdsource workers to interact on these platforms. This allows researchers access to metadata, which is usually proprietary and inaccessible for the typical social media platforms. Furthermore, this allows them to construct traditional controlled experiments with interventions. For example, Liu et al. [29] construct an artificial social platform with recruited MTurk workers and construct an influence ground truth (notably not a full ranking). Guilbeault et al. [30] use a similar approach to investigate opinion polarization. These approaches suffer from the artificial nature of the experiment, and their limitations are typically acknowledged within [29]. Our work does not leverage a synthetic world, and it uses human workers to compare pairs of targets—therefore, the conclusions suffer less from artificially introduced biases. In the second theme, works enlist the knowledge of crowdsource workers and the communities they are situated to identify micro-influencers (often as practical solutions to influence maximization problems). For example, Arous et al. [31] build a framework to ask workers open-ended questions for identifying micro-influencers. However, these approaches address a distinct task (classification rather than ranking), leverage complex worker responses (i.e., open-ended questions), and are constructed predominantly for marketing applications. Our work leverages an easily quantifiable human decision (pairwise comparison) and an algorithmic method to create a ranking of large populations of targets.

*Crowdsource Design.* Several papers point to the importance of properly designing the interface of crowdsource annotation tools and providing the proper context [32, 33]. The design choices directly link to the crowdsourced annotation quality, and prior work suggests keeping tasks simple and clear [34]. Our work defines a measure of performance for the MTurk interface design—the average accuracy of workers—and uses it to optimize the interface design by ablating over several features and procedures.

*Crowdsourcing platforms*, such as Amazon Mechanical Turk (MTurk), allow large pools of human workers to complete tasks that computers cannot do. The research community uses them for labeling tasks (say, identifying objects in pictures) and psychometric studies. They are significantly cheaper, quicker, and induce a more diverse participant pool than traditional surveys [35]. MTurk provides a programmatic interface to serve tasks and process worker responses, making it amenable to active learning setups [36]—i.e., construct the next task based on previous answers. Additionally, researchers have meticulous control of the worker interface and can filter workers by their characteristics.

*Pairwise comparisons.* There are two main psychometric approaches to measurement. The first approach asks people to rate items according to the measured trait (e.g., using a likert scale [37]). However, this direct ranking has scale calibration issues when performed by non-experts [38]. The second approach is the pairwise comparison, which features several advantages; it leads to lower measurement errors than the direct ranking of a set of targets [39]. It induces a quicker and more straightforward experimental task [40], suitable for non-experts in crowdsourcing setups [41] and is amenable to the active learning paradigm [36], which reduces the number of comparisons required for high measurement fidelity. Particularly in noisy and uncertain setups, pairwise comparisons outperform rating scales [42]. Methods for recovering a scoring from pairwise comparisons, such as Bradley-Terry and Thurstonian models, have a strong precedence in psychometric analysis [43, 44].

### 2.2 Pairwise decisions and ranking

The *Pairwise Comparison Matrix* (PCM) $M \in \mathbb{R}^{n\times n}$ represents the outcome of $n$ items being compared where $M[i,j]$ is the number of times item $i$ is favoured to item $j$—denoted hereafter as $j \prec i$.

*Stochastic pairwise decisions and ranking items.* An influence ranking is an ordered list $i \prec j \prec k \prec \cdots \prec z$, meaning that $i$ is the least influential, $j$ is more influential than $i$ but less than $k$, and so on. Going from pairwise comparisons to ranking depends on the difficulty of the pairwise decision task. For simple tasks, human decisions can be considered deterministic—i.e., the same decision is made at multiple repetitions. For such deterministic pairwise decisions, the optimal ranking complexity requires $O(n\log(n))$ comparisons [45]. However, estimating influence is a difficult problem, even for humans; when presented with the same task, two humans might make different choices—e.g., one would say that $i \prec j$ and the other that $j \prec i$. We denote this as a *stochastic pairwise comparison*. When pairwise comparisons are stochastic, all pairs must be compared $t$ times to overcome intransitivity. This requires a dense PCM, which takes $O(tn^2)$ comparisons and is prohibitively expensive when worker remuneration is per comparison.

*Bradley-Terry Model (BT)* [46] proposes a method for ranking individuals with sparse PCM—when only incomplete pairwise comparisons are available. It is commonly used in sports analysis [47, 48] (e.g., ranking chess players from matches) and psychometric studies [43]. The BT model is intimately linked to the work of the Weber-Fechner laws [49], and Thurstone's Law of Comparative Judgement [50]. Each individual $i$ (i.e., a *target*) is ranked by its latent intensity $\theta_i \in \mathbb{R}$. The probability that target $j$ is preferred to $i$ is $\mathbb{P}(i \prec j) = \frac{1}{1+e^{-(\theta_j - \theta_i)}}$. The maximum-likelihood estimates (MLE) $\hat{\boldsymbol{\theta}}$ are computable even from incomplete sets of pairwise comparisons containing circular comparison results (e.g. $i \prec j \prec k \prec i$) [11, 44]; particularly relevant for crowdsourcing experiments where workers make difficult choices differently. Furthermore, adaptive methods for choosing the pairs to compare were proposed [51] to obtain high fidelity measurements with minimal comparisons.

## 3 The empirical influence ranking model

This section introduces the empirical influence methodology—our cost-effective method to construct empirical influence rankings using peer perceptions by MTurk workers. First, we describe the active learning approach that leverages an augmented BT model [11] and the ranking inference procedure (Sect. 3.1). Next, we propose a set of simulation and fitting tools to estimate the required annotation budget (Sect. 3.2). Furthermore, we show the connection between parameters (for modeling systematic noise) and MTurk worker accuracy (Sect. 3.2).

### 3.1 Empirical influence measurements

Building the dense pairwise comparisons matrix $M$ requires $O(tn^2)$ comparisons for $n$ targets (because decisions are stochastic, see Sect. 2.2). As $n$ grows, the process becomes prohibitive using crowdsourcing platforms, where costs are directly proportional to the number of comparisons. The Bradley-Terry model has been successfully applied on sparse versions of $M$ (i.e., not all pairs are compared) to build approximate rankings [52]. The question is selecting which pairs to compare to maximize the ranking quality with the minimum number of comparisons. Passive techniques choose pairs before running the experiment; however, they do not use the information learned during the experiment. Here, we employ a solution that exploits active learning to choose comparisons on the fly. Past comparisons inform future choices, which, in turn, are more informative than random choices.

*Sparse pairwise comparisons matrix via human-in-the-loop active learning.* Our empirical influence quantification method builds on the active learning approach introduced by Maystre and Grossglauser [11]. We use the Quicksort (QS) algorithm to select pairs in the sparse pairwise comparison matrix $M$. We implement a human-in-the-loop system, in which human judges make the pairwise comparisons (using the MTurk platform), while the algorithm chooses which pairs to compare and builds the final ranking. The QS algorithm chooses a pivot point (target) and compares every other target in the set to this pivot. Crowdsourcing workers perform these comparisons. Two partitions are then formed based on these comparisons. Furthermore, we implement QS recursively; at each iteration, the sorting of the left ($<$) and right ($\geq$) subpartitions are performed in parallel, taking full advantage of MTurk's massive worker pool. In its design, QS exploits information from past comparisons to reduce the number of future comparisons required to complete the task, minimizing the total experiment cost.

We compare a pair of targets at most once during each QS execution (denoted as a *run*); usually, multiple runs are required. Maystre and Grossglauser [11] show that Kendall's

Tau—a ranking quality metric – improves with the number of comparisons made, and the estimated ranking approaches asymptotically the true ranking.

*The augmented BT model.* Response fidelity in psychometric experiments suffers from two types of noise. The first type is *systematic noise*, associated with worker subjectivity, worker inauthenticity, and perception biases. This type of noise can be minimized via experimental design interventions (see Sect. 5). The second type of noise is *stochastic noise* that we average out using repeated trials. To account for response fidelity, we use an augmented BT model [11] that introduces the noise $\lambda$ into the probability of preferring the target $j$ over $i$ as

$$\mathbb{P}_{aug}(i \prec j) = \frac{1}{1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}}. \tag{1}$$

Maystre and Grossglauser [11] show that Kendall's Tau deteriorates as noise increases.

*Target ranking inference.* Finally, given an observed set of comparisons $\{i \prec j\}$, we infer the influence scores $\hat{\boldsymbol{\theta}}$ by maximizing the log-likelihood:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{\{i \prec j\}} \log\bigl(\mathbb{P}_{aug}(i \prec j)\bigr) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{\{i \prec j\}} \log\left(\frac{1}{1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}}\right) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} - \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}\bigr) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}\bigr).
\end{aligned} \tag{2}$$

In this work, we solve Eq. (2) using maximum likelihood estimation (MLE) via iterative Luce Spectral Ranking [53]. The following section shows how to estimate the required budget needed to run a real-world experiment.

### 3.2 Estimate required budget: noise, accuracy and simulations

This section introduces tools to estimate the required budget to run an influence estimation experiment at scale. First, we discuss the intertwining of noise, budget, and ranking quality. Next, we introduce a tool to simulate worker decisions with a given accuracy, and we show how to compute the required budget given the expected worker accuracy and the desired ranking quality. Finally, we show the theoretical link between the noise parameter $\lambda$ and the average worker decision accuracy.

*Intertwining of noise, budget, and ranking quality.* Deploying the empirical influence measurement introduced in Sect. 3.1 requires making a tradeoff between three intertwined factors: the noise parameter $\lambda$, budget (number of comparisons which translate into dollars), and ranking quality. For example, a higher noise level would require a higher budget to obtain a ranking of a given quality. Conversely, reducing the systematic noise reduces the required budget. We measure the ranking quality as the Spearman rank correlation between the targets' inferred and real ranking.

*Estimate budget requirements.* It is often desirable to be able to estimate the required budget prior to launching the MTurk experiment. We achieve this in two steps.

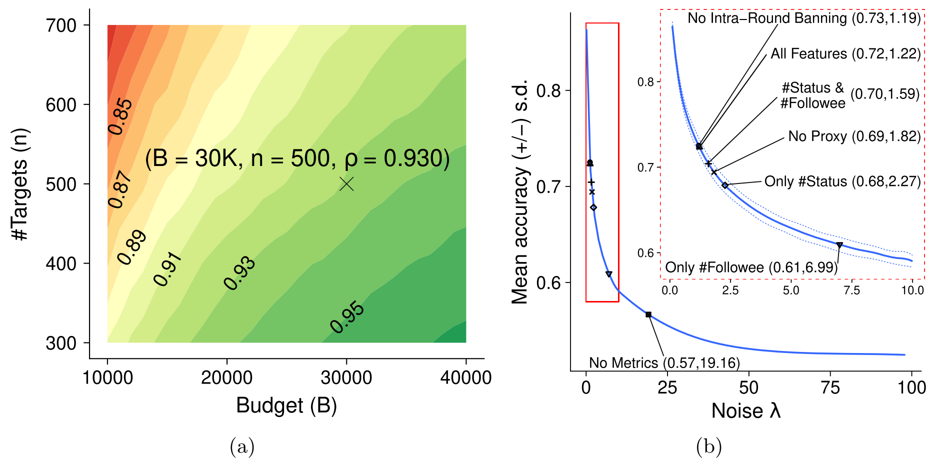


**Figure 2** (a) *Estimate required MTurk budget.* Here we infer the budget, given a required quality of ranking and number of targets. We vary the number of targets ($n$, y-axis) and maximum budget ($B$, x-axis). The color map and contour annotations show the Spearman correlation between the BT-estimated influence ranking ($\hat{\boldsymbol{\theta}}$) and a synthetic ground truth ($\boldsymbol{\theta}$). The cross denotes the chosen setup and estimated budget for our real-world experiments. (b) *Select worker interface design features.* Here we determine design features associated with higher worker decision accuracies, and find the relationship between noise and accuracy. The blue line shows the relation between the average accuracy (y-axis) and the noise ($\lambda$) (averaged over 100 simulations, $n = 500$ targets, $B = 30{,}000$). The points show ablations of design features (see Table 1) and their MLE fitted noise (relevant area zoomed in the inset). As more features are shown, worker accuracy increases

In the *first step*, we simulate the worker's decision process with a given noise parameter, infer the synthetic influence ranking, and compute the ranking quality. The simulation requires three parameters: the maximum number of comparisons $B$ (*budget*), the number of targets $n$ (*#targets*), and the noise parameter $\lambda$ (*noise*). We sample the synthetic latent influence intensities $\theta_i$ from a power-law distribution. We chose power-law as the literature observes that social metrics tend to follow a rich-get-richer paradigm [54]. For example, Twitter follower count is power-law distributed with exponent 2.016 [55], which we use for sampling the synthetic $\theta_i$. For a pair of targets $(i, j)$, our simulated workers produce correct decisions with probability $\mathbb{P}_{aug}(i \prec j)$ (see Eq. (1)), which is completely defined by $\theta_i$, $\theta_j$, and $\lambda$. We use the QS procedure to select $B$ comparisons and compute the BT estimates $\hat{\boldsymbol{\theta}}$ from the recorded responses. Finally, we measure ranking quality $\rho$ as the Spearman correlation between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. Therefore, the first step expresses the ranking quality as $\rho = \text{function}(B, n, \lambda)$.

In the *second step*, we perform a grid search over the budget $B$ and the number of targets $n$. We therefore obtain $B = \text{function}(\rho, n|\lambda)$. We note that $\lambda$ is not a parameter; it is linked to the worker accuracy (see later in this section) and depends on the MTurk interface design (see Sect. 5). Figure 2(a) shows as a colormap $\rho$ as a function of $B$ (x-axis) and $n$ (y-axis). Visibly, for a target quality (the colored area in Fig. 2(a)), the required budget increases with the number of targets. Here, $\lambda = 1.22$ based on our pilot study in Sect. 5. The labeled crossmark in Fig. 2(a) shows the configuration that we use in our experiments in Sect. 5; we chose a correlation level of 0.93 and 500 targets. Therefore, we estimate we require 30,000 comparisons (approximately US$120).

*Link between the noise parameter and worker decision accuracy.* The systematic noise $\lambda$ is intuitively linked to the accuracy of worker decisions: a higher noise is linked to a lower accuracy and vice-versa. Here, we show the theoretical connection between these two quantities. Let the decision of a worker comparing targets $i$ and $j$ be described by a Bernoulli random variable

$$
\begin{aligned}
& X_{ij} \sim Bern\big(\mathbb{P}_{aug}(i \prec j)\big) \\
\text{with } & \mathbb{E}[X_{ij}] = 1 \times \mathbb{P}_{aug}(i \prec j) + 0 \times \big(1 - \mathbb{P}_{aug}(i \prec j)\big) \\
& \quad = \mathbb{P}_{aug}(i \prec j).
\end{aligned}
$$

The MTurk experiment is characterized by a series of Bernoulli trials, one trial per pair $(i,j)$. Consequently, the accuracy of the human choices in the MTurk experiment is $\frac{\sum_{(i,j)} X_{ij}}{N}$. Note that $X_{ij}$ are independent but not identically distributed since they depend on the choice of $(i,j)$ for a given $\lambda$. The expected accuracy of the MTurk experiment over all worker choices is

$$
\mathbb{E}[accuracy] = \mathbb{E}\left[\frac{\sum_{(i,j)} X_{ij}}{N}\right] = \frac{\sum_{(i,j)} \mathbb{E}[X_{ij}]}{N} = \frac{\sum_{(i,j)} \mathbb{P}_{aug}(i \prec j)}{N} \tag{3}
$$

where $N$ is the total number of compared $(i,j)$ pairs. Eq. (3) links the mean worker accuracy and the $\lambda$ noise parameter (via Eq. (1)). Visibly,

$$
\begin{aligned}
\lim_{\lambda \to \inf} \mathbb{E}[accuracy] &= \frac{\sum_{(i,j)} \lim_{\lambda \to \inf} \mathbb{P}_{aug}(i \prec j)}{N} \\
&\overset{\text{cf. Sect. 3.1}}{=} \frac{\sum_{(i,j)} \frac{1}{2}}{N} = \frac{1}{2},
\end{aligned} \tag{4}
$$

i.e., this is the accuracy of unbiased random choice.

We use the synthetic worker decision generator described above to compute the relation between the noise $\lambda$ and the mean accuracy. Figure 2(b) plots this relationship and the standard deviation determined from 1000 process simuations. We make two observations. First, the mean accuracy converges asymptotically to 0.5 as the $\lambda$ increases, as indicated by Eq. (4). Second, we notice that the standard deviations (dotted lines) are minimal, implying the relationship between accuracy and noise is fairly robust. In Sect. 5 we perform a series of pilot experiments to optimize the MTurk worker interface. We observe that, as we refine the worker interface design, the slider on the noise-accuracy line moves towards higher accuracy and lower systematic noise.

## 4 Dataset, implementation, and setup

This section introduces the foundational implementation details for running our QS ranking using real-life crowdsourcing workers. Firstly, we introduce the base experimental setup. Next, we describe the X/Twitter dataset that we use in this work. Finally, we describe how we sample the target and proxy users from the dataset.

*MTurk experimental setup.* The base experimental setup detailed here is used consistently across most variants explored in Sects. 5 and 6. We use the ubiquitous MTurk crowdsourcing platform to implement the QS active learning procedure. The implementation runs QS partitions concurrently, so comparison pairs enter a First-In-First-Out (FIFO) queue and are served to MTurk workers in batches of 10. The workers were presented with two target users and a proxy user (see Table 1 and Sect. 5). Through a pool

**Table 1** Design features for the MTurk user study

| Feature name | Feature description |
|---|---|
| Username, picture & link | Twitter user profile information |
| Description | User-reported description |
| Tweet samples | Sample of 5 tweets emitted by the user |
| Follower Count | The number of people who follow the user |
| Followee Count | The number of people the user follows |
| Status Count | The number of posts the user has authored |
| Proxy User | A third user on whom the influence of each of the targets is projected |
| Qualifications | A mechanism to blocklist low-quality workers |

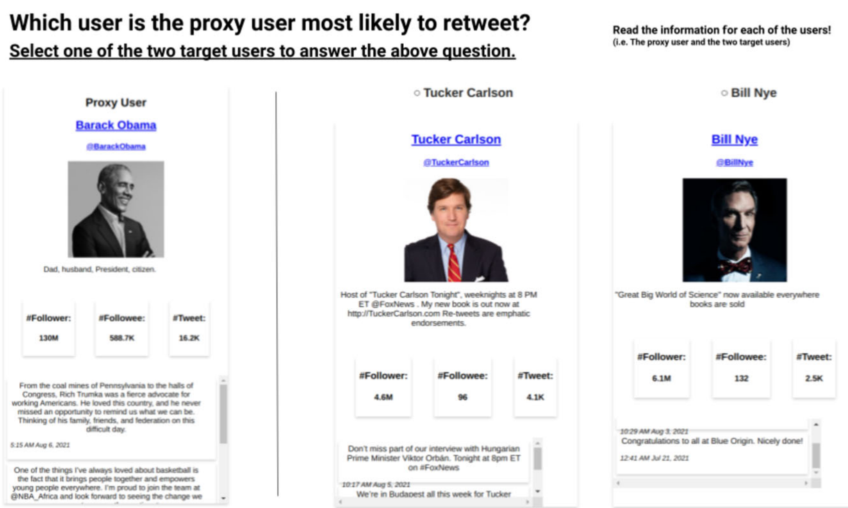


**Figure 3** An example of the user panel within the MTurk interface, showing profile features (name, picture, and handle), metrics (follower, followee, and status counts), and a scrollable sample of authored tweets

of differently worded questions, workers were asked to determine which target user was more influential to the proxy. These questions are *"Which user is the proxy user most likely to retweet?"*, *"Who will the proxy user be more socially influenced by?"*, and *"Which user would sway the proxy user's opinion more?"*. Figure 3 shows the MTurk user interface as it would be presented to a worker.

*#ArsonEmergency Dataset* was collected by Graham and Keller [56] from X/Twitter in the context of the Australian 'Black Summer' Bushfires. It contains discussions around claims that arsonists caused the bushfires—now debunked as misinformation. The dataset was collected between 22 November 2019 and 9 January 2020—using keywords like *arsonemergency*, *bushfireaustralia*, *bushfirecrisis*, and other—, and contains 197,475 tweets emitted by 129,778 users.

*Target and proxy sample selection.* We selected from the #ArsonEmergency dataset a sample of 500 targets and 500 proxies, controlled for availability, language, and Hawkes-modeled influence [57]. The users' availability (suspended or protected status at the time of the experiment) was queried through the Twitter API before the experiment. We selected solely English-speaking users, so the majoritively English-speaking MTurk workers could appropriately judge them. We use a triple-agreement approach between three language detection systems `langid` [58], `cld3` [59], and `whatthelang`. There may be a significant interaction between language and influence, which could affect our sampling.

To remedy this, we verified that the (55.8% remaining) filtered users were uniformly distributed with respect to Hawkes-modeled influence, via a chi-square test at a 95% significance level. From this set of valid users, we used an inverse CDF sampling method, with nearest-matching, to sample users with respect to Hawkes-modeled influence.

## 5 Pilot study: optimize MTurk interface design

In this section, we show how to design the MTurk worker interface to increase the worker decision accuracy and reduce the systematic noise. Intuitively, a proper experimental design and a worker interface presenting the appropriate information increase the accuracy of the decision. We begin by generally describing our pilot study methodology. Next, we detail each design feature and its corresponding impact on decision accuracy. Finally, we apply our design learnings and generate our final empirical measurement for influence.

*Methodology.* An ablative pilot study has a few ingredients: a set of design features, a procedure for running pilots, and a method to compare them. To begin, the set of experimental design features includes the *user component*, *proxies*, and *qualifications*. Table 1 shows the complete list and the feature description. Next, each pilot is a variation of the basic setup with design features added or removed. We performed three QS runs for each pilot and cleared the worker blocklist (see 'qualification system' below) between pilots. For clarity, workers who have participated in a prior pilot could participate in future pilots. In effect, workers may be exposed to different information associated with the same targets between pilots. Given the number of targets and potential comparisons, we assume this *memory bias* is negligible. Furthermore, we apply this protocol because while MTurk worker pools are vast, we do not wish to disenfranchise high-quality workers. Finally, to compare pilots we infer the decision accuracy they induce in workers. We fit the noise hyper-parameter $\lambda$ by minimizing $\sum_i^N (\hat{\theta}_i - \theta_i)^2$, where $\hat{\boldsymbol{\theta}}$ is the influence determined using our QS active learning procedure and $\boldsymbol{\theta}$ is the ground truth. Note that $\lambda$ is a hyper-parameter of our empirical influence measurement model, as it depends on the quality of workers' decisions and not the evaluated targets (see Sects. 3.1 and 3.2). Consequently, $\lambda$ and $\boldsymbol{\theta}$ cannot be jointly fit from pairwise comparisons (see the Appendix for a formal proof). Our pilot study uses follower count as a proxy for $\boldsymbol{\theta}$ because it is a widely adopted metric of influence (although it has been shown to be sub-optimal [54, 60, 61]). Figure 2(b) plots the obtained noise $\lambda$ and corresponding worker accuracy for each ablated MTurk design. In the rest of this section, we detail the impact of each design feature.

*The user component* allows workers to quickly glean the relevant information about users, including the users' names, pictures, descriptions, hyperlinks to their Twitter profile, and a small sample of their tweets presented. We found that the most important user metrics are the *follower count*, *followee count*, and *statuses count*. Figure 2(b) shows that removing these metrics significantly reduces worker accuracy—when we remove all metrics (the `No Metrics` annotated point on Fig. 2(b)), we observe the worst decision accuracy of 0.57. Showing the followee count or the status count significantly increases the accuracy to 0.61 (`Only #Followee`) and 0.68 (`Only #Status`), respectively; when showing both metrics above, the accuracy increases to 0.70 (`#Status & #Followee`). Adding the follower count (`All Features`) further boosts the accuracy to 0.72. The above results suggest that all metrics are independently important signals of influence.

*Proxy users* are used to reduce the effect of a worker's opinion on their influence judgment. In judging between two targets, we ask workers "who the proxy would find more influential". Proxy users do not eliminate the workers' subjectivity; however, they increase the worker decision accuracy by 3%—`No Proxy` has an accuracy of 0.69 compared to `All Features`.

*The qualification system* restricts designated workers from completing tasks. The incentive structure of MTurk encourages workers to do as many tasks as quickly as possible. This leads to workers performing low-quality work; increasing the payment for each HIT (individual piece of work) does not alleviate the problem [62]. We broadly label such workers as *low-quality workers* and implement a blocklist mechanism to stop them from doing additional work and further reduce the decision accuracy.

Within a run, we measure the quality of a worker as the accuracy of the workers' responses concerning the target follower count percentiles. As the task is inherently subjective and difficult, we add some leniency to this banning scheme. Firstly, only comparisons where the difference of follower count percentiles is more significant than 20% are included in determining accuracy. The intuition is that targets with a clear difference in the number of followers should be easier to judge. Secondly, banning is only implemented after completing 100 comparisons that satisfy the prior condition. Lastly, the banning decision is only made once per run (but a banned worker stays banned for all remaining runs). This banning scheme is lenient enough to complete work quickly while restrictive enough that response quality remains high. Counterintuitively, Fig. 2(b) shows that removing the *intra-banning* mechanism (`No Intra-Round Banning`) leads to slightly better performance (accuracy 0.73) than `All Features` (accuracy 0.72). We believe the 0.73 score is too optimistic because low-quality workers were consistently discouraged from our task by the prior implementation of banning; when we removed it to measure impact, they did not return. This also corroborates with previous findings in the literature [62, 63].

*Optimal MTurk interface and required budget.* The above pilot studies lead us to the final MTurk worker interface with `All Features`, shown in Fig. 3, with a worker accuracy of 0.72 (corresponding a systematic noise $\lambda = 1.22$). Given the 500 targets to rank (see Sect. 4) and a desired ranking quality $\rho = 0.930$, we use the budget estimation procedure in Sect. 3.2. We determined that we require 36,252 comparisons. According to MTurk ethics and US Federal minimum wage, we pay MTurk workers US$0.04 per HIT (a HIT contains 10 pairwise comparisons). As a result, we estimate the total cost of building the influence ranking using our proposed QS ranking for 500 targets to US$145. By comparison, building the dense comparison matrix for 500 targets would cost more than US$7000 (see Sect. 2.2).

## 6 Empirical social influence and social cognition

Social influence is difficult to quantify directly; however, several works [14, 15] have shown it is conceptually linked to social cognition. In this section, we illustrate how our empirical influence measure captures the relationship between social influence and the *Big Two* of social cognition [12]; however, the more widely used follower count fails to do so.

*Empirical influence and follower count.* Many prior works use ad hoc proxies for social influence that are simply correlated with it. One example of a widely used measure correlated with social influence is follower count [64–66]; however, it has been shown to be a biased measure of influence [54, 60, 61]. The QS ranking proposed in this work moves towards measuring social influence at scale through peer perception measurements. For example, Fig. 4(a) shows only a moderate correlation between the follower count and our

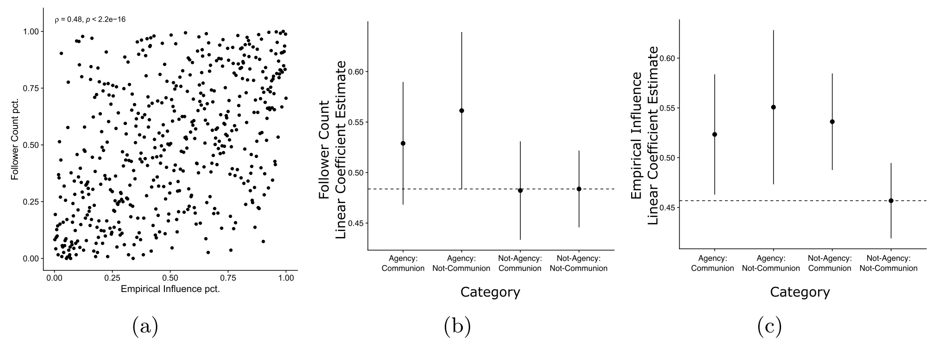

**Figure 4** Plot (a) illustrates the insufficiency of follower count (i.e., the canonical metric) in representing empirical influence. It shows a moderate (spearman) correlation of 0.48 between empirical influence (x-axis) and follower count (y-axis). Plot (b) and (c) illustrate the nessecity of an empirical influence measure, in uncovering relationships with psychosocial attributes, over ad hoc metrics such as follower count. They show the coefficients of a linear regression predicting follower count (b) and empirical influence (c), respectively. The predictors are the use (or lack of use) of agency and communion in user tweets, and there is no intercept. The dashed line shows the estimate for the reference group (using neither agency nor communion), and the range shows a 95% confidence interval. Accordingly, coefficient ranges encompassing the dashed line are not significantly different from the reference group

empirical social influence measure, suggesting a significant variation is not explained by the former. The remainder of this section shows why our empirical social influence estimation is a superior quantification of true social influence by connecting it with social cognition.

*Linking the Big Two of social cognition to influence.* Many factors impact the social influence phenomenon [21], some of which relate to the context of interaction and others to the individuals. Factors related to individuals include likeability and authority (i.e., expertise, confidence, social class, and status). The *Big Two* of social cognition—i.e., agency and communion—are categories of social motives that have been linked to gender [67], and social class [68]. Agency—the drive *to get ahead*, acquire skills, status, and power—is associated with masculinity and independence. Communion—the drive *to get along*, build trust, generate goodwill, and stoke mutual interest—is associated with femininity and dependence. The *Big Two* are intimately related to one's perception of self, others, and groups [13]. Intuitively, social cognition is linked to influence formation (for example, through agentic authority and communal likeability), and prior literature draws links between these social cognition concepts and social influence. For example, Abele and Wojciszke [13] show that liking is related to communion and respect is related to agency (and both liking and respect are features of influence [21]). Furthermore, Tveleneva et al. [14] link communion to higher susceptibility to influence. Frimer et al. [16] find that influential figures (such as public figures) use agentic and communic motives in specific ways. Finally, Abele and Wojciszke [12], Abele and Bruckmüller [69] suggest the Big Two are the fundamental ways we judge others; accordingly, if any personal traits affect influence, it would be these.

*Linking empirical influence and the Big Two.* Here, we analyze the relationship between the *Big Two* and social influence and illustrate the advantage of our empirical measure over ad hoc metrics like the follower count. We quantify the Big Two using a dictionary-based method [70], which matches the n-grams associated with agency and communion with the n-grams in the posts of our 500 target users (see details in Sect. 4). For example, posts containing the n-grams *authoritative* and *persistent* indicate higher agency, while those containing *benevolent* and *kindness* indicate higher communion.[1]

Next, we fit two linear regression models on the same dataset of 500 users: one for the follower count percentile and another for the empirical influence score, using the presence or absence of agentic and/or communic n-grams as predictors. Figures 4(b) and 4(c) show the coefficients for each fitted model. For the follower count (Fig. 4(b)), we observe that no predictor accounting for Big Two features is significantly different from the reference group, as all confidence intervals intersect the dashed line of the reference group (see caption of Fig. 4). However, in Fig. 4(c), we observe that all predictors are significantly different from the reference group. This means using agentic or communal language is associated with higher empirical social influence but not with a higher follower count. In other words, the follower count seems disconnected from social cognition measures, whereas our empirical influence measure is tightly dependent. Furthermore, we observe that agency alone is most highly associated with social influence, followed by communion alone. We speculate this is because X/Twitter is a debating environment [71] where relationships are transient and non-ongoing. In such environments, the agency is more important to project competence and influence quickly. We might expect communion to have a higher role in influence formation on platforms more conducive to forming social groups (such as Facebook). This is an avenue for future work. Interestingly, having both agency and communion leads only to a modest increase in social influence. This is because Abele and Wojciszke [13] show that agency and communion are negatively correlated—showing one dimension makes people assume the opposite about the other. As a result, Gebauer et al. [72] suggest that congruence (showing both dimensions together) leads to ambivalence in an observer. Testing the above hypotheses requires a larger sample, which we leave to future work.

## 7 Discussion

This work's contribution is four-fold; a human-in-the-loop empirical influence measurement framework, simulation and fitting tools for the framework, an empirical study of experimental design context features, and an analysis linking the Big Two of social cognition to social influence. The empirical influence measurement methodology is a novel contribution; robust to noise (see Fig. 2(b)) and highly correlated with true influence in a broad range of flexible simulation studies (see Fig. 2(a)). Furthermore, we find that upon applying the framework to online users, the empirical influence scores are better correlated with factors intuitively related to social influence than the baseline. It is important to note that the empirical method is limited to measuring the peer-perceived social influence of target individuals, which might be epistemologically distinct from latent social influence. We assume that workers are capable of distinguishing the relative social influence between individuals, and this is an accepted sociometric approach [8, 9]. The method could be appropriated to measure other psychosocial attributes of online individuals (such as reputation and trustworthiness) inexpensively, reliably and simply.

[1] See https://osf.io/jfct2 for the complete list of Big Two n-grams.

## Appendix: Bradley-Terry noise invariance

In this section, we show that within the augmented Bradley-Terry model, $\lambda$ and $\boldsymbol{\theta}$ cannot be jointly fit from pairwise comparisons. We show that any ML estimate of $\lambda$ and $\boldsymbol{\theta}$ would not be unique.

Suppose there exists unique ML estimates, $\hat{\lambda}$ and $\hat{\boldsymbol{\theta}}$. Then, consider two constructed quantities, with arbitrary scalar $k$;

$$\lambda^* = k\hat{\lambda},$$
$$\boldsymbol{\theta}^* = k\theta.$$

The log-likelihood for the augmented BT model, is given as

$$\ell(\boldsymbol{\theta},\lambda) = \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(\theta_j-\theta_i)}{\lambda}}\bigr).$$

Consider the maximum log-likelihood,

$$\begin{aligned}
\ell(\hat{\boldsymbol{\theta}},\hat{\lambda}) &= \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(\hat{\theta}_j-\hat{\theta}_i)}{\hat{\lambda}}}\bigr) \\
&= \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(k\theta_j^*-k\theta_i^*)}{k\lambda^*}}\bigr) \\
&= \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-k(\theta_j^*-\theta_i^*)}{k\lambda^*}}\bigr) \\
&= \sum_{\{i \prec j\}} \log\bigl(1 + e^{\frac{-(\theta_j^*-\theta_i^*)}{\lambda^*}}\bigr) \\
&= \ell\bigl(\boldsymbol{\theta}^*,\hat{\lambda}^*\bigr).
\end{aligned}$$

This implies that $\boldsymbol{\theta}^*$ and $\hat{\lambda}$ are ML estimates, which is a contradiction. Therefore unique ML estimates for both $\lambda$ and $\boldsymbol{\theta}$, do not exist.

**Abbreviations**
API, Application Programming Interface; BT, Bradley-Terry; CDF, Cumulative Distribution Function; FIFO, First-in-first-out; HIT, Human Intelligence Task; MLE, Maximum Likelihood Estimation; MTurk, Amazon Mechanical Turk; PCM, Pairwise Comparison Matrix; QS, Quicksort.

**Acknowledgements**
There are no acknowledgments for this paper.

**Author contributions**
Both authors contributed substantially to the manuscript's conceptualization, design, data collection, analyses, interpretation, and drafting. All authors read and approved the final manuscript.

**Funding**
This research was supported by an Australian Government Research Training Program (RTP) Scholarship, the Australian Department of Home Affairs, the Defence Science and Technology Group, the Defence Innovation Network and the Australian Academy of Science. This work has been partially supported by the National Science Centre, Poland (Project No. 2021/41/B/HS6/02798).

**Data availability**
The datasets analyzed are available in the GitHub repository, https://github.com/rohitram96/empirical_influence_measurement. The influence ranking and all code required to reproduce this paper, including the MTurk empirical influence measurement, are included. Auxiliary code and/or data is available from the corresponding author upon reasonable request.

## Declarations

**Competing interests**
The authors declare no competing interests.

Received: 11 March 2024 Accepted: 29 July 2024 Published online: 05 August 2024

**References**

1. Raven BH (1964) Social Influence and Power. University of California, Department of Psychology, California (US)
2. Moussaïd M, Kämmer JE, Analytis PP, Neth H (2013) Social influence and the collective dynamics of opinion formation. PLoS ONE 8(11):78433
3. Tunçgenç B, El Zein M, Sulik J, Newson M, Zhao Y, Dezecache G, Deroy O (2021) Social influence matters: we follow pandemic guidelines most when our close circle does. Br J Psychol 112(3):763–780
4. Schüz B, Papadakis T, Ferguson S (2018) Situation-specific social norms as mediators of social influence on snacking. Health Psychol 37(2):153–159
5. Peng S, Zhou Y, Cao L, Yu S, Niu J, Jia W (2018) Influence analysis in social networks: a survey. J Netw Comput Appl 106:17–32
6. Mason WA, Conrey FR, Smith ER (2007) Situating social influence processes: dynamic, multidirectional flows of influence within social networks. Personal Soc Psychol Rev 11(3):279–300
7. Cork A, Everson R, Levine M, Koschate M (2020) Using computational techniques to study social influence online. Group Process Intergroup Relat 23(6):808–826
8. Prinstein MJ (2007) Assessment of adolescents' preference-and reputation-based peer status using sociometric experts. Merrill-Palmer Q 53(2):243–261
9. Košir K, Pečjak S (2005) Sociometry as a method for investigating peer relationships: what does it actually measure? Educ Res 47(1):127–144
10. Cillessen AH, Marks PE (2011) Conceptualizing and measuring popularity. In: Popularity in the peer system, pp 25–56
11. Maystre L, Grossglauser M (2017) Just sort it! A simple and effective approach to active preference learning. In: ICML
12. Abele AE, Wojciszke B (2013) The Big Two in social judgment and behavior
13. Abele AE, Wojciszke B (2014) Communal and agentic content in social cognition: a dual perspective model. In: Advances in experimental social psychology
14. Tveleneva A, Scholz C, Falk EB, Yoon C, Lieberman MD, Cooper N, O'Donnell MB, Cascio CN (2023) The relationship between agency, communion, and neural processes associated with conforming to social influence. Pers Individ Differ 213:112299
15. Marszał-Wiśniewska M, Siembab M (2012) Power and the self-ascription of agency and communion. Curr Psychol 31:6–16
16. Frimer JA, Walker LJ, Lee BH, Riches A, Dunlop WL (2012) Hierarchical integration of agency and communion: a study of influential moral figures. J Pers 80(4):1117–1145
17. Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1:143–186
18. Macy MW, Ma M, Tabin DR, Gao J, Szymanski BK (2021) Polarization and tipping points. Proc Natl Acad Sci USA 118(50):e2102144118
19. Van de Rijt A (2019) Self-correcting dynamics in social influence processes. Am J Sociol 124:1468–1495
20. Gestefeld M, Lorenz J (2023) Calibrating an opinion dynamics model to empirical opinion distributions and transitions. J Artif Soc Soc Simul 26(4):1–9
21. Cialdini RB (2001) Influence: science and practice. Pearson
22. Cillessen AH (2009) Sociometric methods
23. Zhang Y, Zhang J, Lease M, Gwizdka J (2014) Multidimensional relevance modeling via psychometrics and crowdsourcing. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval
24. Chen C, Eickhoff C (2022) Evaluating search explainability with psychometrics and crowdsourcing. arXiv preprint. arXiv:2210.09430
25. Wetherell EM (2019) The use of crowdsourcing in the development of measurement instruments. PhD thesis, the University of Iowa
26. Milgram S, Gudehus C (1978) Obedience to authority. Ziff-Davis Publishing Company
27. Bhattacharya S, Spann B, Agarwal N (2024) Solidarity to storming: Assessing the socio-technical factors behind modern social movements
28. Shaik M, Yousefi N, Agarwal N, Spann B (2023) Evaluating role of Instagram's multimedia in connective action leveraging diffusion of innovation and cognitive mobilization theories: Brazilian and Peruvian social unrest case studies. In: BESC
29. Liu R, Greene KT, Liu R, Mandic M, Valentino BA, Vosoughi S, Subrahmanian VS (2021) Using impression data to improve models of online social influence. Nature Scientific reports
30. Guilbeault D, Baronchelli A, Centola D (2021) Experimental evidence for scale-induced category convergence across populations. Nat Commun 12(1):327
31. Arous I, Yang J, Khayati M, Cudré-Mauroux P (2020) Opencrowd: a human-AI collaborative approach for finding social influencers via open-ended answers aggregation. In: WWW
32. Finnerty A, Kucherbaev P, Tranquillini S, Convertino G (2013) Keep it simple: reward and task design in crowdsourcing. In: Proceedings of the biannual conference of the Italian chapter of SIGCHI
33. Rahmanian B, Davis JG (2014) User interface design for crowdsourcing systems. In: Proceedings of the 2014 international working conference on advanced visual interfaces
34. Alelyani T, Grogan PT, Tausczik Y, Yang Y (2020) Software crowdsourcing design: an experiment on the relationship between task design and crowdsourcing performance. In: HCI international 2020–late breaking papers: interaction, knowledge and social media: 22nd HCI international conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, proceedings 22

35. Buhrmester M, Kwang T, Gosling SD (2016) Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality data? Perspect Psychol Sci 6(1):3–5
36. Ailon N (2008) Reconciling real scores with binary comparisons: a new logistic based model for ranking. NIPS
37. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol
38. Tsukida K, Gupta MR (2011) How to analyze paired comparison data. Technical report, Washington Univ Seattle Dept of Electrical Engineering
39. Shah NB, Balakrishnan S, Bradley J, Parekh A, Ramchandran K, Wainwright MJ (2016) Estimation from pairwise comparisons: sharp minimax bounds with topology dependence. JMLR
40. Clark AP, Howard KL, Woods AT, Penton-Voak IS, Neumann C (2018) Why rate when you could compare? Using the "elochoice" package to assess pairwise comparisons of perceived physical strength. PLoS ONE 13:e0190393
41. Perez-Ortiz M, Mantiuk RK (2017) A practical guide and software for analysing pairwise comparison experiments. arXiv preprint. arXiv:1712.03686
42. Peysakhovich A, Chiraphadhanakul V, Bailey M (2015) Pairwise choice as a simple and robust method for inferring ranking data. In: WWW 2015 conference proceedings
43. Mair P (2018) Modern psychometrics with R
44. Cattelan M (2012) Models for paired comparison data: a review with emphasis on dependent data. Stat Sci 27:412–433
45. Hoare CA (1962) Quicksort. Comput J
46. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39:324–345
47. Zermelo E (1929) Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. Math Z 29:436–460
48. Cattelan M, Varin C, Firth D (2013) Dynamic Bradley–Terry modelling of sports tournaments. J R Stat Soc, Ser C, Appl Stat 62:135–150
49. Fechner GT (1948) Elements of psychophysics
50. Thurstone LL (1954) The measurement of values. Psychol Rev
51. Settles B (2009) Active learning literature survey
52. Lenton R (2006) Using the method of paired comparisons in non-designed experiments. PhD thesis, Griffith University
53. Maystre L, Grossglauser M (2015) Fast and accurate inference of Plackett-Luce models. NIPS
54. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. In: WSDM
55. Mishra S, Rizoiu M-A, Xie L (2016) Feature driven and point process approaches for popularity prediction. In: CIKM
56. Graham T, Keller TR (2020) Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight. Conversation 10
57. Rizoiu M-A, Graham T, Zhang R, Zhang Y, Ackland R, Xie L (2018) # DebateNight: the role and influence of socialbots on Twitter during the 1st 2016 US presidential debate. In: ICWSM
58. Lui M, Baldwin T (2012) Langid.py: an off-the-shelf language identification tool. In: Proceedings of the ACL 2012 system demonstrations
59. Google: CLD3 (2020)
60. Smith ST, Kao EK, Shah DC, Simek O, Rubin DB (2018) Influence estimation on social media networks using causal inference. In: IEEE Ssp
61. Romero DM, Galuba W, Asur S, Huberman BA (2011) Influence and passivity in social media. In: Ecml Pkdd
62. Yuan L, Rizoiu M-A (2022) Detect hate speech in unseen domains using multi-task learning: a case study of political public figures. http://arxiv.org/abs/2208.10598
63. Law A (2021) Exposing the Stance of Reddit Users Towards Brexit. PhD thesis, The Australian National University. https://doi.org/10.13140/RG.2.2.32779.98083
64. Cha M, Haddadi H, Benevenuto F, Gummadi K (2010) Measuring user influence in Twitter: the million follower fallacy. In: ICWSM
65. Frantz TL, Cataldo M, Carley KM (2009) Robustness of centrality measures under uncertainty: examining the role of network topology. Comput Math Organ Theory 15(4):303–328
66. Riddell J, Brown A, Kovic I, Jauregui J (2017) Who are the most influential emergency physicians on Twitter? West J Emerg Med 18(2):281–287
67. Hsu N, Badura KL, Newman DA, Speach MEP (2021) Gender, "masculinity, " and "femininity": a meta-analytic review of gender differences in agency and communion. Psychol Bull 147(10):987–1011
68. Aydin AL, Ullrich J, Siem B, Locke KD, Shnabel N (2019) The effect of social class on agency and communion: reconciling identity-based and rank-based perspectives. Soc Psychol Pers Sci 10(6):735–745
69. Abele AE, Bruckmüller S (2013) The big two of agency and communion in language and communication. In: Social cognition and communication
70. Pietraszkiewicz A, Formanowicz M, Gustafsson Sendén M, Boyd RL, Sikström S, Sczesny S (2019) The big two dictionaries: capturing agency and communion in natural language. Eur J Soc Psychol 49:871–887
71. Calderon P, Ram R, Rizoiu M-A (2024) Opinion market model: stemming far-right opinion spread using positive interventions. In: Proceedings of the international AAAI conference on web and social media, pp 1–12. http://arxiv.org/abs/2208.06620
72. Gebauer JE, Maio GR, Pakizeh A (2013) Feeling torn when everything seems right: semantic incongruence causes felt ambivalence. Pers Soc Psychol Bull 39(6):777–791


## Publisher's Note