Incentive-Compatible Classification[*]

Yakov Babichenko, Oren Dean, Moshe Tennenholtz
Technion — Israel Institute of Technology
Haifa, Israel

Abstract

We investigate the possibility of an incentive-compatible (IC, a.k.a. strategy-proof) mechanism for the classification of agents in a network according to their reviews of each other. In the α-classification problem we are interested in selecting the top α fraction of users. We give upper bounds (impossibilities) and lower bounds (mechanisms) on the worst-case coincidence between the classification of an IC mechanism and the ideal α-classification.

We prove bounds which depend on α and on the maximal number of reviews given by a single agent, Δ. Our results show that it is harder to find a good mechanism when α is smaller and Δ is larger. In particular, if Δ is unbounded, then the best mechanism is trivial (that is, it does not take into account the reviews). On the other hand, when Δ is sublinear in the number of agents, we give a simple, natural mechanism, with a coincidence ratio of α.

1 Introduction

There are many situations in which peer agents have binary, directed interactions with each other, and in which one side can rate his experience from the interaction, or rather rate the other agent. The following are just a few examples:

1. E-commerce sites in which buyers might also be sellers (e.g., ebay.com, amazon.com).
2. Academic paper reviewers for a conference might themselves be authors of papers submitted to the same conference.
3. Employees in an organisation are sometime asked to fill out a sociometric overview of their fellow friends.

In all of the above examples, a coordinator/manager is classifying the agents in the system according to the reviews they received. In the e-commerce example, the top-rated sellers will appear higher and more frequently in search results; the academic conference will only accept a certain top-rated portion of the papers. Employees with higher sociometric results have better chances at a promotion. A natural problem arises—in order to maximize one's relative rating, it is a dominant strategy in these situations to give a harsh critique to all interactions. In this paper, we model the agents and their interactions as a directed network and ask whether it is possible to offer an incentive-compatible (IC) mechanism to select a subset which represents the top-rated ("worthy") agents. We measure the quality of a mechanism as the resemblance between the selected set of the mechanism and the set of top-rated agents, in the worst case. We investigate the relation between the quality of the best possible mechanism to two parameters: (i) the maximal number of reviews a single agent can issue (the maximal out-degree in the network, denoted Δ), and (ii) the fastidiousness of the system (the relative size of the selected set, denoted α).

Our contribution

In this paper we investigate the existence of an α-classification IC mechanism in weighted networks. Weighted networks without any limitation were not considered before as a framework for selection mechanisms ((Kurokawa et al. 2015) considered only Δ-regular weighted networks; the assumption of Δ-regularity somewhat simplifies the optimisation criterion since in this case the optimisation for average in-weights is equivalent to the optimisation for the sum of in-weights.). The most significant novelty of our model is the consideration of mechanisms which classify the agents to "worthy" and "unworthy", in contrast to previous works which only considered k-subset selection (k-selection) mechanisms. Our optimisation criterion is the coincidence between the mechanism's classification and the ideal classification. The difference from k-selection is two-fold. First, we do not know the exact size of the set which we need to select. Second, we try to select as many of the right (truly worthy) agents and not the wrong agents regardless of how high their in-degree is; this is very different from k-selection, which just looks for a subset with high in-degree (even if it is completely disjoint to the optimal subset).

We prove upper and lower bounds on the quality of the best possible IC classification mechanism. We chart the behaviour of these bounds as a function of Δ and α. We show that as Δ grows (agents are allowed to review many others), the possibility of an IC classification mechanism narrows down until for large values of Δ the best mechanism is one of the two trivial mechanisms: select all the agents as wor-

[*]This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^{o} 740435).

thy, or select every agent independently with probability 1/2. We show the reverse behaviour with α: as we lower α (the system is more picky about its worthy-classified agents), the quality of the best possible IC classification mechanism decays to zero.

On the other hand, for fixed α and for Δ which is negligible with respect to the number of agents, we provide a mechanism with a positive quality. The idea behind this mechanism is based on a well-known practice to partition the agents into three subsets: absolutely worthy, borderline, and absolutely unworthy. Unlike the well-known practice, our mechanism suggests to classify an agent into these categories after ignoring his reviews on others. This makes the mechanism IC, but complicates the performance analysis of the mechanism.

Previous works (e.g. (Alon et al. 2011), (Kurokawa et al. 2015)) showed the existence of an optimal k-selection mechanisms when k is large (say $k = \omega(1)$ or $k = \omega(\Delta)$ with regard to the number of agents). As explained above, these mechanisms only select a k-subset of agents with high in-degree, while an α-classification mechanism needs to select as many of worthy agents and not unworthy agents. This extra predicament shows in the results as we bound the quality of any IC mechanism away from 1 (i.e., for any $\alpha < 1$ there is no ideal IC classification mechanism). This also shows the significance of our mechanism which, under reasonable assumptions, selects not only *good* agents but the *right* agents.

The graphs in Figure 1 summarize our main findings. Graph (a) shows our bounds on the quality of any IC mechanism as a function of α when Δ is negligible with respect to the number of agents. The grey dashed line is the ideal mechanism. The upper bound (red line) shows that for any $\alpha < 1$, there is a gap between the best possible mechanism and the ideal mechanism, and this disparity is larger for lower values of α. The blue line denotes the quality of our proposed mechanism, and so the green area is what we know about the possible value of the quality of the best IC mechanism. Graph (b) shows the quality of any IC mechanism as a function of α when Δ is not bounded. Here the upper and lower bounds coincide to a single line, that is, we know what is the best possible quality for any value of α.

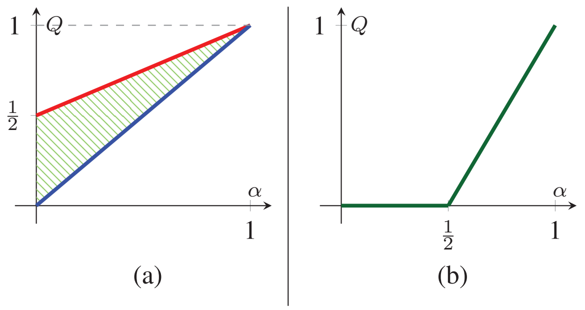

Figure 1: Our bounds for the quality of an IC classification mechanism as a function of α. In (a) $\Delta = o(n)$. The red/blue graphs denote our upper/lower bounds and the green area denotes the known feasible quality range. In (b) $\Delta = n - 1$. The single green graph is the quality of any IC mechanism.

Related work

There have been in the past decade quite a few works on IC (sometimes called strategyproof or impartial) selection mechanisms in networks. The most similar model to ours is (Kurokawa et al. 2015). In that paper the authors considered the k-selection in a weighted Δ-regular network (that is, every agent has Δ out-edges and Δ in-edges), where k and Δ are constants. They considered IC mechanisms which try to optimize the average in-weights of the selected set. Their main result is a probabilistic mechanism which is optimal when Δ is negligible with respect to k.

Other papers on this subject only consider unweighted networks, which can be seen as a special case of the weighted networks with weights in $\{0, 1\}$. These works can be divided into two flavours:

(a) *Optimisation works* which try to optimize the total in-degree of the selected agent or set of agents. Examples of this group are Alon et al.; Fischer and Klimm; Bjelde, Fischer, and Klimm; Bousquet, Norin, and Vetta (2011; 2014; 2017; 2014). Another example is Babichenko, Dean, and Tennenholtz (2018) in which the authors try to optimize the progeny of the selected agent. These works differ in several parameters of the problem, such as deterministic/probabilistic mechanisms; exact selection (selected set must be of size k) or inexact selection (selected set is of size at most k); and the sub-family of networks considered (all networks/m-regular networks/acyclic networks/etc.).

(b) *Axiomatic works* which define a set of axioms and investigate the possibility/impossibility of mechanisms that fulfil maximal subsets of these axioms. Examples of this group are Holzman and Moulin; Mackenzie; Aziz et al. (2013; 2015; 2016). In (Altman and Tennenholtz 2008) the authors considered the possibility of complete ranking mechanisms under certain axioms.

Paper organization. In Section 2, we formally present our model and main results. In Section 3, we prove our impossibility (upper bound) propositions (Propositions 10 and 11). These proofs rely on an extension of our model to a symmetric, probabilistic mechanism; this extension is formally defined in the beginning of Section 3. In Section 4, we present and prove a mechanism with a positive quality when the number of reviews of a single agent is negligible with respect to the number of agents (Proposition 12). In Section 5, we conclude and discuss our results.

2 Model and main results

Let $N = [n]$ be a set of n agents.[1] We represent the interactions between the agents as a directed graph, $G(N, E)$; thus an edge (x, y) means that agent x interacted with agent y and is allowed to review this agent. Let $E_{in}(x), E_{out}(x)$ be the sets of in-edges and out-edges of x, respectively. We assume that each agent in the network is in control of the weights of

[1] We assume n is large as we are generally interested in results which are asymptotic in n. For the same reason we habitually drop floors and ceilings for easier reading.

his outgoing edges. These weights, which are real numbers in the interval [-1,1],[2] represent the reviews of the source agent for the target agents. Thus, the reviews of agent $x \in N$ for his interactions are $\{w_e | e \in E_{out}(x)\}$. After all agents submitted reviews on their interactions,[3] we get a weighted, directed graph; from now on we assume all the edges of G are weighted.

Based on the reviews that agent x received, $\{w_e | e \in E_{in}(x)\}$, we define his score and his relative ranking in the system. The *score* of agent x is the average of weights on $E_{in}(x)$:

$$s(x, G) = \begin{cases} \dfrac{\sum_{e \in E_{in}(x)} w_e}{|E_{in}(x)|}, & E_{in}(x) \neq \emptyset, \\ 0, & E_{in}(x) = \emptyset. \end{cases} \tag{1}$$

The *ranking* of agent x is the number of agents who strictly have a better score than him: $r(x, G) = |\{y \in N | s(y, G) > s(x, G)\}|$.[4] For a given real parameter $\alpha \in (0, 1)$ we consider as *worthy* the subset of agents who are in the top α-ranking:[5]

$$I_\alpha(G) = \{x \in N | r(x) < \alpha n \}.$$

Notice that the size of $I_\alpha(G)$ is at least αn, but might be higher in case of ties. For instance, in the empty graph, $I_\alpha(G(N, \emptyset)) = N$ for all α. We denote by $\overline{I_\alpha}(G) = N \backslash I_\alpha(G)$, the subset of *unworthy* agents.

Our goal is to offer an IC mechanism which selects a set which is as similar as possible to the subset of worthy agents. Formally, let $\mathcal{G}(n)$ be the family of all [-1,1]-weighted, directed networks on n nodes and let $P(N)$ be the power set of N.

Definition 1. *A* classification mechanism *is a function* $M : \mathcal{G}(n) \to P(N)$.

The set $M(G)$ is the subset of agents which the mechanism M classifies as worthy in the network G. We denote by $\overline{M}(G) = N \backslash M(G)$ the subset classified as unworthy by the mechanism. Notice that the definition of a mechanism depends on n through its dependence on $\mathcal{G}(n)$ and N. We abuse notation and regard a single mechanism M as if it represents a series of mechanisms—one for every natural n.

The IC requirement means that an agent's classification is not influenced by his own reviews; that is, changing the weights on the out-edges of an agent does not alter his classification.

Definition 2. *A classification mechanism M is* incentive-compatible *if for every $n \in \mathbb{N}, G, G' \in \mathcal{G}(n), x \in N$, such that: $E(G) = E(G')$ and $\forall e \in E \backslash E_{out}(x)$, $w_e(G) = w_e(G')$,*

$$x \in M(G) \iff x \in M(G').$$

To define a measure for the quality of the mechanism, we first define a measure of coincidence between $M(G)$ and $I_\alpha(G)$:

$$\mathcal{C}(M(G), I_\alpha(G)) =$$
$$\frac{1}{n} \sum_{x \in N} \begin{cases} 1, & x \in (M(G) \cap I_\alpha(G)) \cup (\overline{M}(G) \cap \overline{I_\alpha}(G)), \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

In other words, $\mathcal{C}(M(G), I_\alpha(G)$ gives one point for every agent that $M(G)$ classified correctly and takes one point for every agent which was classified erroneously; the result is normalised by the number of agents.[6] This measure can be somewhat simplified to get,[7]

$$\begin{aligned} &\mathcal{C}(M(G), I_\alpha(G)) = \\ &\frac{1}{n}(|((M(G) \cap I_\alpha(G)) \cup (\overline{M}(G) \cap \overline{I_\alpha}(G)))| - |M(G) \oplus I_\alpha(G)|) \\ &= \frac{1}{n}(|N \backslash (M(G) \oplus I_\alpha(G))| - |M(G) \oplus I_\alpha(G)|) \\ &= 1 - \frac{2}{n}|M(G) \oplus I_\alpha(G)|. \end{aligned}$$

Our main theorems imply that the possibility of an IC mechanism which guarantees a fixed level of coincidence depends on two parameters. The first is α. The second is the maximal out-degree in the network (i.e., the maximal reviews an agent can issue), denoted by Δ. Intuitively, if Δ is high, then an unworthy agent might use his influence on the score of Δ worthy agents to improve his ranking and be considered worthy. If α is relatively low, then these manipulations might influence a large portion (or all) of the worthy-classified agents, which makes it harder to find an IC mechanism with high coincidence measure. Our results will show that this intuition is indeed correct. Let $\mathcal{G}(n, \Delta)$ be the family of all networks on n nodes with maximal out-degree Δ. The quality of a mechanism for given α, Δ, is the limit when n goes to infinity of the worst case of the coincidence measure:

$$Q_{\alpha,\Delta}(M) = \lim_{n \to \infty} \min_{G \in \mathcal{G}(n,\Delta)} \mathcal{C}(M(G), I_\alpha(G)). \tag{3}$$

We will prove upper and lower bounds on $Q_{\alpha,\Delta}$ for any IC classification mechanism.

Main results

We start by defining two trivial IC mechanisms. 'Trivial' means that they classify the nodes without any regard to the edges' weights.

Let M_N be the complete mechanism, i.e., $M_N(G) = N$ for all $G \in \mathcal{G}(n)$.

Proposition 3. *For any α, Δ, $Q_{\alpha,\Delta}(M_N) = 2\alpha - 1$.*

[2] A review of '0' is the neutral review. For example, an agent with all-zero in-edges is rated in the same way as an agent with no in-edges (see the definition of $s(\cdot)$ in (1)).

[3] We can set a zero-weight on all interactions which were not reviewed (see Footnote 2), hence we may assume w.l.o.g. that all interactions have been reviewed.

[4] From now on, when G is clear from the context, we just write $s(x), r(x)$.

[5] We think of α as the selectiveness of the system: a lower value for α means that the tag of being worthy is more prestigious.

[6] Other measures might be appropriate for different applications. We discuss two of these alternatives in Section 5. The main conclusions from our results stay the same in these variations.

[7] The operator $\oplus$ is the "exclusive or".

Proof. Since $|N \oplus I_\alpha(G)| = |\overline{I_\alpha}(G)| = n - |I_\alpha(G)| \le n(1-\alpha)$, we get that for any graph,

$$\mathcal{C}(M_N(G), I_\alpha(G)) = 1 - \frac{2}{n}|N \oplus I_\alpha(G)| \ge 1 - \frac{2n(1-\alpha)}{n} = 2\alpha - 1. \quad \square$$

Our second trivial mechanism, $M_{1/2}$, selects every node to be worthy with probability 1/2, independently. We use here the concept of a probabilistic mechanism intuitively; in the beginning of Section 3, we formally extend our model to include this kind of mechanism. Mechanism $M_{1/2}$ correctly classifies nodes x with probability 1/2, hence $Q_{\alpha,\Delta}(M_{1/2}) = 0$ for all α, Δ.

Our first result is a strong impossibility, saying that when Δ is large, one of the two trivial mechanisms is the best possible.

Theorem 4.

(a) For $\alpha \ge \frac{1}{2}$ *and* $\Delta \ge 2(1-\alpha)n$, $Q_{\alpha,\Delta} = 2\alpha - 1$.

(b) For $\alpha < \frac{1}{2}$ *and* $\Delta = (1-o(1))n$, $Q_{\alpha,\Delta} = 0$.

Proof. This is a direct consequence of Proposition 3, our observation for $M_{1/2}$ and Proposition 11. □

Our second result is a non-trivial, yet quite natural, mechanism with a better quality than the complete mechanism, provided $\Delta = o(n)$. The idea behind this mechanism is to recognize three subsets of agents: absolutely worthy, absolutely unworthy, and borderline. The first two subsets contain those agents who will not be able to change their classification, no matter what their reviews will be. The fact that Δ is negligible guarantees that the absolutely worthy and absolutely unworthy sets will include a large portion of the true worthy and unworthy subsets, respectively. The mechanism then classifies the absolutely worthy agents as worthy and the absolutely unworthy as unworthy. If we allow the mechanism to be probabilistic, we can select each of the borderline agents to be worthy with probability 1/2; this strategy assures that these agents will not hurt the quality (but will not help it either). We also provide a deterministic version of this mechanism which always classifies correctly almost half of the borderline agents. The following theorem summarises our knowledge when $\Delta = o(n)$.

Theorem 5. *For any* α *and* $\Delta = o(n)$, $\alpha \le Q_{\alpha,\Delta} \le \frac{1}{2}(1+\alpha)$.

Proof. The lower bound comes from an analysis of the above-described probabilistic mechanism; see Proposition 13. We also prove the existence of a deterministic version of this mechanism in Proposition 12. The upper bound is proved in Proposition 10. □

3 Impossibilities

As promised, we start by extending our model to include probabilistic mechanisms. A probabilistic mechanism assigns each node a probability of being worthy.

Definition 6. *A* probabilistic classification mechanism *is a function* $M_p : N \times \mathcal{G}(n) \to [0,1]$.

To get a concrete selected set from a probabilistic mechanism, we select each node independently with his assigned probability. In other words, the probability of subset $X \subseteq N$ to be selected under mechanism M_p in the graph G is

$$\Pr[X|M_p(G)] = \prod_{x\in X} M_p(x,G) \prod_{x\notin X}(1 - M_p(x,G)). \tag{4}$$

The IC requirement translates to the requirement that an agent cannot influence his own selection probability.

Definition 7. *A probabilistic classification mechanism* M_p *is* incentive-compatible *if for every* $n \in \mathbb{N}, G, G' \in \mathcal{G}(n), x \in N$*, such that:* $E(G) = E(G')$ *and* $\forall e \in E\backslash E_{out}(x)$, $w_e(G) = w_e(G')$, $\qquad M_p(x,G) = M_p(x,G')$.

The coincidence of $M_p(G)$ with $I_\alpha(G)$ is naturally extended using expectation over the selected set:

$$\mathcal{C}(M_p(G), I_\alpha(G)) = \sum_{X\in P(N)} \Pr[X|M_p(G)]\mathcal{C}(X, I_\alpha(G))$$

$$= \frac{1}{n}\sum_{X\in P(N)} \Pr[X|M_p(G)]$$

$$\cdot \sum_{x\in N} \begin{cases} 1, & x \in (X \cap I_\alpha(G)) \cup (\overline{X} \cap \overline{I_\alpha}(G)) \\ -1, & \text{otherwise.} \end{cases}$$

Changing the summation order and inserting (4) we get,

$$\mathcal{C}(M_p(G), I_\alpha(G)) =$$

$$\frac{1}{n}\sum_{x\in N}\sum_{X\in P(N\backslash\{x\})}\prod_{y\in N\backslash\{x\}} \begin{cases} M_p(y,G), & y \in X \\ 1 - M_p(y,G), & y \notin X \end{cases}$$

$$\cdot \begin{cases} M_p(x,G) - (1 - M_p(x,G)), & x \in I_\alpha(G) \\ (1 - M_p(x,G)) - M_p(x,G), & x \in \overline{I_\alpha}(G). \end{cases}$$

Since $\sum_{X\in P(N\backslash\{x\})}\prod_{y\in N\backslash\{x\}} \begin{cases} M_p(y,G), & y \in X \\ 1 - M_p(y,G), & y \notin X \end{cases} = 1$,

$$\mathcal{C}(M_p(G), I_\alpha(G)) = \frac{1}{n}\sum_{x\in N}(2M_p(x,G) - 1) \cdot \begin{cases} 1, & x \in I_\alpha(G) \\ -1, & x \in \overline{I_\alpha}(G) \end{cases} \tag{5}$$

$$= \frac{\overline{I_\alpha}(G) - I_\alpha(G)}{n} + \frac{2}{n}\sum_{x\in N} M_p(x,G) \cdot \begin{cases} 1, & x \in I_\alpha(G) \\ -1, & x \in \overline{I_\alpha}(G). \end{cases} \tag{6}$$

The quality of a probabilistic mechanism can now be defined exactly as in (3):

$$Q_{\alpha,\Delta}(M_p) = \lim_{n\to\infty} \min_{G\in\mathcal{G}(n,\Delta)} \mathcal{C}(M_p(G), I_\alpha(G)).$$

From Definitions 6 and 7 and the coincidence definition above, it is clear that the IC (deterministic) mechanisms for which we initially defined our problem are a special case of IC probabilistic mechanisms. Hence we may prove our upper bounds (i.e., impossibilities) for probabilistic mechanisms. We furthermore show that it is enough to consider a subfamily of *symmetric*, IC, probabilistic mechanisms.

Definition 8. *A probabilistic mechanism* M_p *is* symmetric *if for any network* $G \in \mathcal{G}(n)$ *and two isomorphic nodes* $x, y \in N$, $M_p(x,G) = M_p(y,G)$.

Claim 9. *Let M_p be any IC, probabilistic, classification mechanism. Then there is an IC, symmetric, probabilistic classification mechanism M'_p with $Q_{\alpha,\Delta}(M'_p) \geq Q_{\alpha,\Delta}(M_p)$.*

Proof. Let $S(N)$ be the set of permutations over N. For $\pi \in S(N)$, let G_π be the graph which is isomorphic to G under the automorphism defined by π. We define the mechanism M'_p:

$$M'_p(x, G) = \frac{1}{n!} \sum_{\pi \in S(N)} M_p(\pi(x), G_\pi). \tag{7}$$

Mechanism M'_p is clearly IC, since M_p is IC and otherwise the calculation above is irrelevant of any of the weights in G. It is also symmetric, since for two isomorphic nodes, x, y, the following two sets of the couples are exactly the same:

$$\{(\pi(x), G_\pi) | \pi \in S(N)\} = \{(\pi(y), G_\pi) | \pi \in S(N)\}.$$

It remains to show that the quality of M'_p is at least that of M_p. Planting (6) into (7) and changing the summation order we get that for any G,

$$\begin{aligned}
&\mathcal{C}(M'_p(G), I_\alpha(G)) = \\
&\frac{\overline{I_\alpha}(G) - I_\alpha(G)}{n} + \frac{2}{n \cdot n!} \sum_{x \in N} \sum_{\pi \in S(N)} M_p(\pi(x), G_\pi) \\
&\qquad\qquad \cdot \begin{cases} 1, & x \in I_\alpha(G_\pi) \\ -1, & x \in \overline{I_\alpha}(G_\pi) \end{cases} \\
=&\frac{\overline{I_\alpha}(G) - I_\alpha(G)}{n} + \frac{1}{n!} \sum_{\pi \in S(N)} \Bigg[\frac{2}{n} \sum_{x \in N} M_p(\pi(x), G_\pi) \\
&\qquad\qquad \cdot \begin{cases} 1, & x \in I_\alpha(G_\pi) \\ -1, & x \in \overline{I_\alpha}(G_\pi) \end{cases} \Bigg]
\end{aligned}$$

Let $Q = Q_{\alpha,\Delta}(M_p)$. For any $\epsilon > 0$, there is n_0 such that for all $n > n_0$ and for all $G \in \mathcal{G}(n, \Delta)$, $\mathcal{C}(M_p(G, I_\alpha(G))) \geq Q - \epsilon$, which means that

$$\frac{2}{n} \sum_{x \in N} M_p(x, G) \cdot \begin{cases} 1, & x \in I_\alpha(G) \\ -1, & x \in \overline{I_\alpha}(G) \end{cases} \geq Q - \epsilon - \frac{\overline{I_\alpha}(G) - I_\alpha(G)}{n}.$$

Hence, $\mathcal{C}(M'_p(G), I_\alpha(G)) \geq Q - \epsilon$. Since this is true for any ϵ, we get that $Q_{\alpha,\Delta}(M'_p) \geq Q$. □

We are now ready to prove our two impossibility propositions which imply the upper bounds of Theorems 4 and 5.

Proposition 10. *For all α, Δ, $Q_{\alpha,\Delta} \leq \frac{1}{2}(1 + \alpha)$.*

Proof. Let v be a distinct node. Partition $N \backslash \{v\}$ into three sets, A, B, C, of sizes $\frac{(1-\alpha)n}{2}, \frac{(1-\alpha)n}{2}, \alpha n - 1$, respectively. Let G be the network in which every node in $A \cup B$ has an out-edge to v, and C is a cycle. Set the weights on the edges from A to v to be 1, the weights on the edges from B to v to be $-1 + \frac{1}{(1-\alpha)n}$, and the weights on C to be 1. For any $a \in A, b \in B, c \in C$ their scores are: $s(c) = 1$, $s(a) = s(b) = 0$, and $s(v) = \frac{|A| + |B|(-1 + 1/(1-\alpha)n)}{|A| + |B|} = \frac{1}{2(1-\alpha)n} > 0$. Hence $I_\alpha(G) = C \cup \{v\}$. Let M be an IC, probabilistic and symmetric classification mechanism. By symmetry, we may denote $M(a, G) = \mu$ for any $a \in A$. Using (5) we get that,

$$\begin{aligned}
\mathcal{C}(M(G), I_\alpha(G)) &\leq \frac{1}{n}((1 - 2\mu)|A| + |B| + |C| + 1) \\
&= 1 - \mu(1 - \alpha).
\end{aligned} \tag{8}$$

Now choose a distinct node $a_0 \in A$ and change the weight on its out-edge to v to be $-1 + \frac{1}{(1-\alpha)n}$; we refer to this network as G'. The score of v has dropped in $\frac{2 - 1/(1-\alpha)n}{|A| + |B|} = \frac{2 - 1/(1-\alpha)n}{(1-\alpha)n} > \frac{1}{2(1-\alpha)n}$, for n large enough. Hence $s(v, G') < 0$. Since the rest of the scores are the same in G and G', we get that $I_\alpha(G') = N \backslash \{v\}$. By IC, $M(a_0, G') = M(a_0, G) = \mu$, and by symmetry $M(b, G') = \mu$ for any $b \in B$. We now get ,

$$\begin{aligned}
\mathcal{C}(M(G'), I_\alpha(G')) &\leq \frac{1}{n}(|A| + (2\mu - 1)|B| + C + 1) \\
&= 1 - (1 - \mu)(1 - \alpha) = \alpha + \mu(1 - \alpha).
\end{aligned} \tag{9}$$

From (8) and (9), $Q_{\alpha,\Delta} \leq \min\{1 - \mu(1 - \alpha), \alpha + \mu(1 - \alpha)\}$. Comparing the two terms to find the optimal value for μ we find that

$$\begin{aligned}
1 - \mu(1 - \alpha) = \alpha + \mu(1 - \alpha) &\iff \mu = \frac{1}{2} \\
\Longrightarrow Q_{\alpha,\Delta} \leq 1 - \frac{1 - \alpha}{2} &= \frac{1 + \alpha}{2}. \quad \square
\end{aligned}$$

Proposition 11. *Let $m = \min\left\{2(1 - \alpha), \frac{\Delta}{n}\right\}$. Then $Q_{\alpha,\Delta} \leq 1 - m$.*

Proof. Consider two cases:

Case I: $m \geq 2\alpha$.
Let $A, B \subseteq N$ be two disjoint subsets of size $nm/2$. Let G be the graph in which $A \cup B$ is a clique and there are no other edges. We set the weights on all the out-edges of nodes in A to be 0, and the weights on all the out-edges of nodes in B to be 1. Hence every $a \in A$ has a score of $s(a) = \frac{|B|}{|A \cup B|}$, every $b \in B$ has a score of $s(b) = \frac{|B| - 1}{|A \cup B|}$ and the score of every $c \notin A \cup B$ is 0. Since $|A| = \frac{mn}{2} \geq \alpha n$, $I_\alpha(G) = A$. Let M be an IC, probabilistic, symmetric mechanism. By the symmetry of M, all the vertices of B get the same probability. Denote it μ. Let b be a distinct vertex in B. Let G' be the graph we get when we nullify all the weights on the outgoing edges of b. Since b is now isomorphic to the vertices in A, we get by IC and the symmetry of M that $\forall a \in A$, $M(a, G') = M(b, G') = M(b, G) = \mu$. We see that there is a trade-off in the value of μ; on the one hand, we need it to be high if we want a good quality on G', but on the other hand, it needs to be low for a good quality on G. The idea of the proof is that for n large enough, we can repeat this process and show that in essence all the vertices in $A \cup B$ in the graph G (or at least in one of the graphs we get from G after a negligible number of steps) should have approximately the same probability. This

implies that

$$Q_{\alpha,\Delta} \leq \frac{1}{n}(|A|(2\mu - 1) + |B|(1 - 2\mu) + |N\backslash(A \cup B)|)$$
$$= \frac{2\mu - 1}{n}(|A| - |B|) + 1 - \frac{|A \cup B|}{n} = 1 - m.$$

To make this claim precise, suppose for contradiction that $Q_{\alpha,\Delta}(M) \geq 1 - m + \epsilon$ for some $\epsilon > 0$. For $k \leq X$, X to be found, we let $A^k, B^k \subseteq N$ be two disjoint subsets with sizes $|A^k| = mn/2 + k$, $|B^k| = mn/2 - k$. Define the graph G^k in which $A^k \cup B^k$ is a clique, and the weights on the outgoing edges of vertices in A^k are all 0, and the weights on the outgoing edges of vertices in B^k are all 1. Notice the following:

a) The vertices in A^k are symmetric, and so are the vertices in B^k.
b) $I_\alpha(G^k) = A^k$.
c) If we nullify the outgoing edges of one of the nodes $b \in B^k$ then we get the graph G^{k+1}in which $b \in A^{k+1}$.

By the symmetry of M, we may denote for any k, $\mu_a^k = M(a, G^k)$ for all $a \in A^k$, and $\mu_b^k = M(b, G^k)$ for all $b \in B^k$. By (c) and IC of M, $\mu_b^k = \mu_a^{k+1}$. By the assumption on the quality of M:

$$\frac{1}{n}(|A^k|(2\mu_a^k - 1) + |B^k|(1 - 2\mu_b^k) + |N\backslash(A^k \cup B^k)|)$$
$$\geq 1 - m + \epsilon,$$
$$\frac{1}{n}((mn/2 + k)(2\mu_a^k - 1) + (mn/2 - k)(1 - 2\mu_b^k) + (1 - m)n)$$
$$\geq 1 - m + \epsilon,$$
$$(mn + 2k)\mu_a^k - (mn - 2k)\mu_b^k - 2k \geq \epsilon n.$$

Summing up this inequality for $0 \leq k \leq X$ and substituting μ_b^k for μ_a^{k+1} we get,

$$mn\mu_a^0 - (mn - 2X)\mu_b^X + 2\sum_{k=1}^{X}(2k - 1)\mu_a^k - X(X + 1) \geq X\epsilon n.$$

Let $\mu_{max} = \max_{1 \leq k \leq X} \mu_a^k$. If $X = o(n)$, then for n large enough $mn \geq 2X$. We strengthen the inequality by removing the negative terms on the left-hand side, and replacing all the μ_a^k with μ_{max}:

$$\mu_{max}\left(mn + 2\sum_{k=1}^{X}(2k - 1)\right) \geq X\epsilon n$$
$$\mu_{max}(mn + 2X^2) \geq X\epsilon n$$
$$\Longrightarrow \mu_{max} \geq \frac{\epsilon X}{m + 2X^2/n}.$$

Now taking $X = 2m/\epsilon$ we get that for n large enough $\mu_{\max} > 1$, which is a contradiction.

<u>*Case II:*</u> $m < 2\alpha$
The proof is very similar. We define three subsets, A, B, C of size $nm/2, nm/2$ and $(\alpha - m/2)n$, respectively.[8] Let G be the graph in which $A \cup B$ is a clique and C is a cycle. Again we set all the weights on the out-edges of the nodes in A to be 0, and weights on all the out-edges of nodes in B to be 1. The edges in the cycle in C also get a weight of 1. Thus now $s(c) = 1 > s(a) > s(b)$ for any $c \in C, a \in A, b \in B$. Since $|A \cup C| = \alpha n$, we have $I_\alpha(G) = A \cup C$. Let M be an IC, probabilistic, symmetric mechanism. Using the same technique as before, we get that all the vertices in $A \cup B$ should get the same probability, which we call μ. Since $I_\alpha(G) = A \cup C$ we can bound as before:

$$Q_{\alpha,\Delta}(M) \leq$$
$$\frac{1}{n}(|A|(2\mu - 1) + |B|(1 - 2\mu) + |N\backslash(A \cup B)|) = 1 - m.$$

The formal argument is precisely the same. □

4 Non-trivial mechanism

In this section we show the existence of a mechanism with quality α, provided $\Delta = o(n)$. Specifically, we will show the following:

Proposition 12. *For any α and any $\Delta \leq \min\{\frac{1}{3}\alpha n, \frac{1}{3}(1-\alpha)n\}$, there is an IC classification mechanism with quality $\alpha - \frac{3\Delta}{n}$.*

For better exposition, we start by presenting a probabilistic mechanism which is slightly better.

Proposition 13. *For any α and any $\Delta \leq \alpha n$, there is an IC probabilistic classification mechanism with quality $\alpha - \frac{\Delta}{n}$.*

For a graph G and $x \in G$, denote by G_x the graph we get when we set all the weights on the out-edges of x to be -1. Let $\beta = \alpha - \frac{\Delta}{n}$. The idea is to partition N into three subsets:

$$W(G) = \{x \in N | x \in I_\beta(G_x)\},$$
$$U(G) = \{x \in N | x \in \overline{I_\alpha}(G_x)\},$$
$$B(G) = N\backslash(W \cup U) = \{x \in N | x \in I_\alpha(G_x)\backslash I_\beta(G_x)\}.$$

Notice that the definitions of W, U, B are IC, in the sense that the sorting of x to one of these sets does not depend on his own reviews (since we fixed all his reviews to -1 before choosing his set). If $x \in W$, then his ranking in G_x is less than βn: $r(x, G_x) < \beta n$. The real reviews of x may increase the score of at most Δ agents, which means that $r(x, G) \leq r(x, G_x) + \Delta < \beta n + \Delta = \alpha n$; hence $x \in I_\alpha(G)$. We have proved that $W \subseteq I_\alpha(G)$. We think of W as the *absolutely worthy* agents. Similarly, the set U is the set of *absolutely unworthy* agents: for $x \in U$, $r(x, G) \geq r(x, G_x) \geq \alpha n$ (increasing his reviews can only hurt his relative ranking), and $U \subseteq \overline{I_\alpha}(G)$. The set B contains all the borderline agents: these are the agents which we are not sure how to classify. We define the following probabilistic mechanism:

$$M_p(x, G) = \begin{cases} 1, & x \in W(G), \\ 0, & x \in U(G), \\ 1/2, & x \in B(G). \end{cases}$$

Since W, U, B are IC, this mechanism is IC. The mechanism is always correct in the classification of the agents in $W \cup U$. We set the probability of agents in B to be 1/2 so that the expected contribution of the agents in B to the coincidence

[8] Since we assume $m < 2\alpha$, $|C| > 0$.

measure is zero. We get that for any $G \in \mathcal{G}(n, \Delta)$ with $\Delta \leq \alpha n$,

$$\mathcal{C}(M_p(G), I_\alpha(G)) = \frac{1}{n}(|W(G)| + |U(G)|) \geq \frac{|W(G)|}{n} \geq \beta = \alpha - \frac{\Delta}{n}.$$

This completes the proof of Proposition 13.

The idea for the deterministic mechanism is similar. This mechanism selects as worthy all the agents in W and as unworthy all the agents in U. For the agents in B we need to find an IC, deterministic way to classify them such that in every network about half of them are rightly classified. Notice first that if $|U| \geq |B| - 2\Delta$ then

$$\mathcal{C}(M(G), I_\alpha(G)) \geq \frac{1}{n}\left(|W| + |U| - |B|\right) \geq \frac{1}{n}(\beta n - 2\Delta) = \alpha - \frac{3\Delta}{n}.$$

Let $\mathcal{L}(B)$ be a linear ordering of B according to $r(\cdot, G)$ and breaking ties lexicographically.[9] Let B^+ be the top half nodes in B according to $\mathcal{L}$. For $x \in B$ we denote by B_x the set B in G_x.[10] Similarly, let $B^+_x = B^+(G_x)$. We now complete our definition of mechanism M by setting for every $x \in B$, $x \in M(G) \iff x \in B^+_x$. That is, $x \in B$ is accepted if and only if it is ranked in the top half of B_x when breaking ties lexicographically. The mechanism does not use the weights on the outgoing edges of x to determine its classification, hence it is IC. We complete the proof of Proposition 12 with the following lemma.

Lemma 14. *Suppose that* $\Delta \leq \frac{1}{3}(1-\alpha)n$. *For any graph* $G \in \mathcal{G}(n, \Delta)$ *with* $|U| < |B| - 2\Delta$ *at least* $\frac{1}{2}(|B \cup U| - \Delta)$ *of the agents in* $B \cup U$ *are classified correctly.*

Using this lemma we get that for any such graph,

$$\begin{aligned}&\mathcal{C}(M(G), I_\alpha(G)) \\ &\geq \frac{1}{n}\left(|W| + \frac{1}{2}(|B \cup U| - \Delta) - \frac{1}{2}(|B \cup U| + \Delta)\right) \\ &\geq \frac{1}{n}\left(|W| - \Delta\right) \geq \frac{1}{n}\left(\beta n - \Delta\right) = \alpha - \frac{2\Delta}{n},\end{aligned}$$

as required.

Proof of Lemma 14. Consider two cases.
Case I: $|I_\alpha(G)| \geq \alpha n + 2\Delta$.
In this case $B \subseteq I_\alpha \backslash I_\beta$, since if $x \notin I_\alpha(G)$, then in G_x there are at least $|I_\alpha| - \Delta \geq \alpha n + \Delta$ agents with a higher score than $s(x)$; hence $x \notin I_\alpha(G_x)$. Moreover, for any $x \in B$, $B_x \subseteq I_\alpha(G)$, since for any $y \notin I_\alpha$, there are in $G_{x,y}$ at least $|I_\alpha| - 2\Delta \geq \alpha n$ agents with a higher score than $s(y)$. Let B^* be the top $\frac{1}{2}(|B| - \Delta)$ ranked agents in B according to $\mathcal{L}(B)$. We claim that the agents in B^* are all classified by our mechanism to be worthy, which is the correct classification (since $B \subset I_\alpha(G)$, as explained). The reason is that when we set the weight on an edge (x, y) with $x \in B^*$ to -1, we add at most one agent to B which is ranked above x (this might happen when $y \in W$), or we remove at most one agent from B which is ranked below x (this might happen when $y \in B$ and is ranked below him). Thus x must be in the top-half ranked agents in B_x. Since all the agents in U are correctly classified, we get that at least $|B^*| + |U| = \frac{1}{2}(|B| + 2|U| - \Delta)$ are classified correctly.
Case II: $|I_\alpha(G)| < \alpha n + 2\Delta$.
Let B_* be the bottom $\frac{1}{2}(|B| - |U|) - \Delta$ ranked agents in B according to $\mathcal{L}(B)$.[11] Since $|B_* \cup U| = \frac{1}{2}(|B| + |U|) - \Delta \leq \frac{1}{2}(1-\beta)n - \Delta = \frac{1}{2}(1-\alpha)n - \frac{1}{2}\Delta$, and $|\overline{I_\alpha}(G)| = n - I_\alpha(G) > (1-\alpha)n - 2\Delta \geq \frac{1}{2}(1-\alpha)n - \frac{1}{2}\Delta$, we get that $B_* \subseteq \overline{I_\alpha}(G)$. It is therefore enough to show that for any $x \in B_*$, $x \notin B^+_x$. Indeed, lowering the weights on the out-edges of x can advance x in the ranking of B by at most Δ places (if the edges are to nodes in B that are ranked higher than x), and it might also add all the nodes of U to $|B|$; in any case x will still be in the bottom half of B_x. □

5 Conclusions

We have introduced a generic model for the classification of agents to worthy and unworthy according to their own reviews of each other. We draw two general conclusions regarding the existence of an IC classification mechanism:

1. If Δ is large, there is no good mechanism in the sense that the best mechanisms do not take into consideration the reviews.
2. If Δ is negligible with respect to the number of agents, there is a mechanism with a positive quality, but not with quality 1.

Our measure for the coincidence between the selected set and the true worthy agents (2) assumes that every classification/misclassification has the same value/price. In other applications it makes more sense to consider only the classification/misclassification of the worthy agents, which leads to the following measure:

$$\mathcal{C}(M(G), I_\alpha(G)) = \frac{1}{|I_\alpha(G)|} \sum_{x \in M(G)} \begin{cases} 1, & x \in I_\alpha(G), \\ -1, & x \notin I_\alpha(G). \end{cases}$$

Yet another measure might consider all the agents but normalise the classification/misclassification of an agent according to his true set:[12]

$$\begin{aligned}&\mathcal{C}(M(G), I_\alpha(G)) = \\ &\frac{|M(G) \cap I_\alpha(G)|}{2|I_\alpha(G)|} + \frac{|\overline{M}(G) \cap \overline{I_\alpha}(G)|}{2|\overline{I_\alpha}(G)|}.\end{aligned}$$

We mention here that using these measures does not change our two conclusions above; in other words, our conclusions are intrinsic in the problem and not the result of a specific measure. The exact claims and proofs can be found in the appendix of the full version; see (Babichenko, Dean, and Tennenholtz 2019).

[9] That is, for $x, y \in B$, $x \succ_{\mathcal{L}} y \iff (r(x) > r(y)) \vee ((r(x) = r(y)) \wedge (x < y))$.

[10] To be clear, for $x, y \in N$, let $G_{x,y}$ be the graph in which we set all the weights on the outgoing edges of x and y to -1. Then $B_x(G) = B(G_x) = \{y \in N | y \in I_\alpha(G_{x,y}) \backslash I_\beta(G_{x,y})\}$.

[11] Notice that $|B_*| > 0$ due to our assumption on $|U|$.

[12] Though here we need to assume that $\overline{I_\alpha}(G)$ is not empty, or define the measure for this case separately.

References

Alon, N.; Fischer, F.; Procaccia, A.; and Tennenholtz, M. 2011. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK XIII, 101–110.

Altman, A., and Tennenholtz, M. 2008. Axiomatic foundations for ranking systems. *Journal of Artificial Intelligence Research* 31(1):473–495.

Aziz, H.; Lev, O.; Mattei, N.; Rosenschein, J. S.; and Walsh, T. 2016. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 390–396.

Babichenko, Y.; Dean, O.; and Tennenholtz, M. 2018. Incentive-compatible diffusion. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 1379–1388.

Babichenko, Y.; Dean, O.; and Tennenholtz, M. 2019. Incentive-compatible classification. https://arxiv.org/abs/1911.08849.

Bjelde, A.; Fischer, F.; and Klimm, M. 2017. Impartial selection and the power of up to two choices. *ACM Trans. Econ. Comput.* 5(4).

Bousquet, N.; Norin, S.; and Vetta, A. 2014. A near-optimal mechanism for impartial selection. In Liu, T.-Y.; Qi, Q.; and Ye, Y., eds., *Web and Internet Economics*, 133–146. Cham: Springer International Publishing.

Fischer, F., and Klimm, M. 2014. Optimal impartial selection. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, 803–820.

Holzman, R., and Moulin, H. 2013. Impartial nominations for a prize. *Econometrica* 81(1):173–196.

Kurokawa, D.; Lev, O.; Morgenstern, J.; and Procaccia, A. D. 2015. Impartial peer review. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 582–588.

Mackenzie, A. 2015. Symmetry and impartial lotteries. *Games and Economic Behavior* 94:15–28.