Corruption Detection on Networks

Noga Alon [*] Elchanan Mossel[†] Robin Pemantle[‡]

May 20, 2015

Abstract

We consider the problem of corruption detection on networks. We show that graph expansion is necessary for corruption detection and discuss algorithms and the computational hardness of the problem.

1 Introduction

In this paper we study the problem of *corruption detection* on networks. Given a network of agents, a subset of whom are corrupt, our goal is to find as many non-corrupt agents as possible. Neighboring vertices audit each other. We assume that truthful (non-corrupt) agents report the status of their neighbors accurately. We make no assumption on the report of corrupt agents. For example, two corrupt neighbors may collude and report each other as non-corrupt. Similarly, a corrupt vertex may prefer to report the status of some its neighbors accurately, hoping that this will establish a truthful record for itself. Moreover, we assume that the corrupt agents may coordinate their actions in an arbitrary fashion.

[*] Sackler School of Mathematics and Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel and School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540. nogaa@tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by an ISF grant, by the Israeli I-Core program and by the Oswald Veblen Fund.

[†] Department of Statistics, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104 and Departments of Statistics and Computer Science, U.C. Berkeley, 367 Evans Hall, Berkeley CA 94720. mossel@wharton.upenn.edu. Research supported by NSF grant CCF 1320105, DOD ONR grant N00014-14-1-0823, and grant 328025 from the Simons Foundation.

[‡] Department of Mathematics, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104, USA, pemantle@math.upenn.edu. Research supported in part by NSF grant # DMS-1209117

The corruption detection problem is motivated by corruption detection in organizations and financial institutions and by network auditing in network security. Furthermore it is a natural generalization of the *machine problem*.

The Machine Testing Machine Puzzle. Consider a factory that produces machines. The machines are used to test other machines. We call a machine *truthful* if it functions properly and corrupt otherwise. Given a batch of a 100 machines, exactly 51 of which are truthful:

- Can you find all of the 51 truthful machines in the batch?
- How can this be achieved using the minimal number of tests?

It is not hard to see that the answer to the first item is yes. The second part of the puzzle is a bit more challenging (and is left to entertain those readers who have not seen the puzzle before). Note that

- The machine problem above is a special case of the corruption detection problem on the complete graph on 100 vertices with exactly 51 truthful agents. However, in this problem we allow adaptive algorithms whereas in our corruption detection problem we consider nonadaptive ones.
- It is clear that if the number of corrupt machines is at least 50, then it is impossible to detect even one truthful machine. For example, in the case where there are 50 machines of each type we may consider the following strategy of corrupt machines. A corrupt machine will report a corrupt machine truthful and truthful machine corrupt. It is clear that in this case, there is no way to distinguish between the corrupt and truthful machines.

  We were not able to locate the source of the puzzle. The problem has a similar flavor to the truth telling and lying natives puzzle. It is also motivated by problems in Byzantine computing.

Byzantine computing, Intrusion Detection and Fault Localization in Computer Networks The problem we study here can be viewed as a model for detection of corrupt agents in a computer network, or corrupt components in on computer. The problem of computing with faulty components have been idealizes in terms of Byzantine computing [8]. This problem is sometimes studied under the name of fault localization (see e.g. [7]). Agents might also be corrupt due to intrusion [11].

While the model presented here does not capture all the aspects of corruption in computer networks, it still provides some insight as to the desired structure of a corruption tolerant networks. Note that our model is weaker than standard computer network models as neighbors only report to the network monitor the truthfulness status of their neighbors. On the other hand, it is also stronger as we assume that the monitor observes all reports even from deep inside a corrupt subnetwork.

Corruption Networks in the Social Sciences The original motivation for our work is corruption detection in social and economic networks. Examples of such networks may include different government agencies in a country, the network of banks in the EU or the network of hospitals in a geographic location. Our goal is to understand which network structures are more amenable to corruption and which are more robust against it.

Social scientists have studied many aspect of corruption networks, see e.g.

[12, 13, 6]. However, to the best of our knowledge, prior to this work there is no systematic study of the effect of the network structure on corruption detection.

1.1 Formal Definitions and Main Results

Consider a network of agents represented by a finite directed graph $G = (V, E)$. Each vertex can be either truthful, or corrupt. We denote by B the set of corrupt agents and by T the set of truthful ones. Thus $V = T \cup B, T \cap B = \emptyset$. For each vertex u and each of its out-neighbors v, u examines v and reports about his type. If u is truthful, he reports the truth, that is, reports that $v \in T$ if indeed $v \in T$ and reports that $v \in B$ if $v \in B$. If $u \in B$, then he reports adversarially, independently of the actual nature of v. We assume that the corrupt vertices can cooperate in an arbitrary fashion. The question we address is under what conditions on the graph G and the number of truthful vertices it is possible to identify at least one truthful vertex, with certainty. It is easy to see that this is impossible if $|T| \leq |B|$. Indeed, if $V = V_1 \cup V_2 \cup W$ is a partition of V into 3 pairwise disjoint sets where $|V_1| = |V_2|$ (and W may be empty), then the corrupt agents can ensure that all the reports in the two scenarios $T = V_1$, $B = V_2 \cup W$ and $T = V_2$, $B = V_1 \cup W$ will be identical. As there is no common truthful agent in these two possibilities, no deterministic algorithm can locate a truthful agent with no error.

The main result is that if the graph is a good bounded degree directed expander, in the sense described below, and we have a majority of truthful agents, it is possible to identify most of the truthful agents, whereas if it is far from being an expander this is impossible even if the number of truthful vertices is guaranteed to be at least $0.99|V|$.

We first consider the case of symmetric directed graphs, that is, graphs in which (u, v) is a directed edge iff (v, u) is such an edge. This case is somewhat simpler, and is equivalent to considering G as an undirected graph in which each vertex reports about all his neighbors.

For a positive $\delta < 1/8$ call a graph $G = (V, E)$ on a set of n vertices a δ good expander if any set U of at most $2\delta n$ vertices has more than $|U|$ neighbors outside U, and there is an edge between any pair of sets of vertices provided one of them is of size at least δn and the other is of size at least $n/4$. Standard results about expanders (see, e.g., [1], Corollary 1) imply that this holds for Ramanujan graphs or random regular graphs with degrees at least c/δ for an appropriately chosen absolute constant c. The main result for the undirected case is the following.

Theorem 1.1 *Let $G = (V, E)$ be a δ-good expander and suppose $V = T \cup B$, $T \cap B = \emptyset$ and $|T| > |B|$. Then when getting the reports of each vertex of G about all its neighbors we can identify a subset $T' \subset T$ and a subset $B' \subset B$ so that $|T' \cup B'| > (1 - \delta)n$. That is, we will be able to recover the type of almost all vertices of G.*

Moreover, if $|T| > (1/2 + \delta)n$ then there is a linear time algorithm that identifies subsets as above from the given reports.

We note that the algorithm in the proof of the theorem is exponential if we only assume that $|T| > |B|$ (or if we assume that $|T| > (1/2 + \mu)n$ for a small fixed $\mu = \mu(\delta)$). The fact that the detection algorithm is not efficient when we only assume that T is just a little bit bigger than B is not a coincidence. Indeed, the algorithm described in the proof of the theorem, presented in the next section, provides a set T of more than $n/2$ truthful agents, which is consistent with the reports obtained, when such a set exists. We show that the problem of producing such a set when it exists is NP-hard, even when restricted to bounded degree expanders (and even if we ensure that there is such a set of size at least $n/2 + \eta n$.)

Theorem 1.2 *For any $\delta > 0$ there exists an $\eta > 0$ such that the following promise problem is NP-hard. The input is a graph $G = (V, E)$ with $|V| = n$, which is a δ-good expander along with the status of u reported by v and vice versa for every edge $e = (u, v) \in E$. The promise is that either*

- *There exists a partition of $V = T \cup B$ which is consistent with all of the reported values and $|T| \geq n/2 + \eta n$, or*

- *All partitions $V = T \cup B$ which are consistent with the reported value satisfy $|T| \leq n/2 - \eta n$.*

The objective is to distinguish between the two options above.

We also prove the following, which shows that expansion is essentially necessary for solving the detection problem.

Theorem 1.3 *Let $G = (V, E)$ be a graph on n vertices so that it is possible to remove at most ϵn vertices of G and get a graph in which any connected component is of size at most ϵn. Then even knowing that $V = T \cup B$ with $T \cap B = \emptyset$ and $|T| \geq (1 - 2\epsilon)n$ there is no deterministic algorithm that identifies even a single member $t \in T$ from the reports of all vertices. In particular, this is the case for planar graphs or graphs with a fixed excluded minor even if $\epsilon = \Theta(n^{-1/3})$.*

1.2 Results for directed graphs

Next we consider directed graphs. This is motivated by the fact that in various auditing situation it is unnatural to allow u to inspect v whenever v inspects u. In fact, it may even be desirable not to allow any short cycles in the directed inspection graph. For a fixed $\delta < 1/16$, call a directed graph $G = (V, E)$ on n vertices a δ-good-directed expander if the following conditions hold.

(i) For any set $U \subset V$ of size at most $4\delta n$, $|N^+(U) - U| > |U|$, where $N^+(U)$ is the set of all out-neighbors of V.

(ii) For any two disjoint sets of vertices A and B so that $|A| \geq \delta n$ and $|B| \geq n/4$ there is at least one directed edge from A to B and at least one directed edge from B to A.

We first show that for any fixed positive $\delta < 1/16$ there are bounded degree δ-good directed expanders which contain no short cycles (even ignoring the orientation of edges).

Lemma 1.4 *There are two absolute positive constants c_1, c_2 so that for any fixed $0 < \delta < 1/16$ there is a constant $d < c_1/\delta$ and infinitely many values of n for which there is a δ-good directed expander on n vertices in which the total degree of each vertex is d and there is no cycle of length smaller than $c_2 \log n / \log d$ (of any orientation).*

We then show that those directed graphs are good for our detection problem.

Theorem 1.5 *Let $G = (V, E)$ be a δ-good directed expander and suppose $V = T \cup B$, $T \cap B = \emptyset$ and $|T| > |B|$. Then when getting the reports of each vertex of G about all its out-neighbors we can identify a subset $T' \subset T$ and a subset $B' \subset B$ so that $|T' \cup B'| > (1-\delta)n$. That is, we will be able to recover the type of almost all vertices of G.*

Moreover, if $|T| > (1/2 + 2\delta)n$ then there is a linear time algorithm that identifies subsets as above from the given reports. If we only assume that $|T| > |B|$ then the detection algorithm is exponential.

2 Proofs

2.1 Undirected graphs

Proof of Theorem 1.1: Let H be the spanning subgraph of G in which a pair of vertices u and v is connected iff u reports that $v \in T$ and v reports that $u \in T$. Let $V_1, V_2, \ldots, V_s$ be the sets of vertices of the connected components of H.

Claim 2.1 *All the vertices of each V_i are of the same type, that is, for each $1 \leq i \leq s$, either $V_i \subset T$ or $V_i \subset B$.*

Proof: Suppose u and v are neighbors in H. If $u \in T$ then $v \in T$ (as u reports so). If $u \in B$, then $v \in B$ (as v reports that $u \in T$). □

Call a component of H truthful if it is a subset of T, else it is a subset of B and we call it corrupt.

Let H' be the induced subgraph of G on the set T of all good vertices.

Claim 2.2 *Any connected component of H' is also a connected component of H.*

Proof: If $u, v \in T$ are adjacent in G (and hence in H'), they are adjacent in H as well, by definition and by the fact that each of them reports honestly about his neighbors. Thus each component C' of H' is contained in a component C of H. However, no $v \in T$ is adjacent in H to a vertex $w \in B$, implying that in fact $C' = C$ and establishing the assertion of the claim. □

Claim 2.3 *The graph H' contains a connected component of size at least $|T| - \delta n > (1/2 - \delta)n$.*

Proof: Assume this is false and the largest connected component of H' is on a set of vertices U_1 of size smaller than $|T| - \delta n$. Since the total number of vertices of H' is $|T| > n/2$, it is easy to check that one can split the connected components of H' into two disjoint sets, each of total size at least δn. However, the bigger among the two is of size bigger than $n/4$, and hence, since G is a δ-good expander, there is an edge of G between the two groups. This is impossible, as it means that there is an edge of G between two distinct connected components of H'. □

The analysis so far allow us to prove the easy part of the theorem.

Claim 2.4 *If $|T| > (1/2+\delta)n$ then there exists a linear time algorithm which finds $T' \subset T$ and $B' \subset B$ such that $|T' \cup B'| \geq (1-\delta)n$.*

Proof: Note that if $|T| > (1/2+\delta)n$ then Claim 2.3 implies that H must contain a connected component of size bigger than $n/2$, which must be truthful. Thus, if this is the case, more than $n/2$ of the truthful vertices of G can be identified by the simple, linear time algorithm that computes the connected components of H. Moreover, since all vertices but at most δn are among their neighbors, this enables us to identify the types of all vertices besides less than δn. □

It remains to show that even if we only assume that $|T| > n/2$ then we can still identify correctly most of the truthful vertices. We proceed with the proof of this stronger statement.

By Claims 2.2 and 2.3 it follows that H contains at least one connected component of size at least $(1/2-\delta)n \geq 3/8n$. If H contains only one such component, then this component must consist of truthful agents, and we can identify all of them. Otherwise, there is another connected component of size at least $(1/2-\delta)n$, and as there is no room for more than two such components, there are exactly two of them, say V_1 and V_2. Note that by the properties of G there are edges of G between V_1 and V_2, and hence it is impossible that both of them are truthful components. As one of them must be truthful, it follows that exactly one of V_1 and V_2 is a truthful component and the other corrupt. We next show that we can identify the types of both components.

Construct an auxiliary weighted graph S on the set of vertices $1, 2, \ldots, s$ representing the connected components $V_1, V_2 \ldots, V_s$ as follows. The weight w_i of i is defined by $w_i = \frac{|V_i|}{|V|}$. Two vertices i and j are connected iff there is at least one edge of G that connects a vertex in V_i with one in V_j. Call an independent set in the graph S *large* if its total weight is bigger than $1/2$. Note that by the discussion above T must be a union of the form $T = \cup_{i \in I} V_i$, where I is a large independent set in the graph S. In order to complete the argument we prove the following.

Claim 2.5 *Either there is no large independent set in S containing 1, or there is no large independent set in S containing 2.*

Proof: Assume this is false, and let I_1 be a large independent set in S containing 1, and I_2 a large independent set in S containing 2. To get a contradiction we show that for $w(I_1) = \sum_{i \in I_1} w_i$ and $w(I_2) = \sum_{i \in I_2} w_i$ we have $w(I_1) + w(I_2) \leq 1$ (and hence it is impossible that each of them has total weight bigger than a half).

To prove the above note, first, that the two vertices 1 and 2 of S are connected (as each corresponds to a set of more than $(1/2 - \delta)n$ vertices of G, hence there are edges of G connecting V_1 and V_2). Therefore I_1 must contain 1 but not 2, and I_2 contains 2 but not 1.

If there are any vertices i of S connected in S both to 1 and to 2, then these vertices belong to neither I_1 nor I_2, as these are independent sets. Similarly, if a vertex i is connected to 1 but not to 2, then it can belong to I_2 but not to I_1, and the symmetric statement holds for vertices connected to 2 but not to 1. So far we have discussed only vertices that can belong to at most one of the two independent sets I_1 and I_2. If this is the case for all the vertices of S, then each of them contributes its weight only to one of the two sets and their total weight would thus be at most 1, implying that it cannot be that the weight of each of them is bigger than 1/2, and completing the proof of the claim. It thus remains to deal with the vertices of S that belong to both I_1 and I_2. Let $J \subset \{3, 4, \ldots, s\}$ be the set of all these vertices. Note, first, that the total weight of the vertices in J is at most 2δ, as the total weight of 1 and 2 is at least $2(1/2 - \delta) = 1 - 2\delta$. Note also that by the discussion above each $j \in J$ is not a neighbor of 1 or of 2. By the assumption about the expander G the total weight of the vertices that are neighbors of vertices in J and do not belong to J is bigger than the total weight of the vertices in J. Indeed, this is the case as the number of neighbors in G of the set $\cup_{j \in J} V_j$ that do not lie in this set is bigger than the size of the set. We thus conclude that if $J' = N_S(J) - J$ is the set of neighbors of J that do not belong to J, then the total weight of the vertices in J' exceeds the total weight of the vertices in J, and the vertices in J' belong to neither I_1 nor I_2. We have thus proved that the sum of weights of the two independent sets I_1 and I_2 satisfies

$$w(I_1) + w(I_2) \leq 2w(J) + (1 - w(J) - w(J')) \leq w(J) + w(J') + (1 - w(J) - w(J')) = 1$$

contradicting the fact that both I_1 and I_2 are large. This completes the proof of the claim. □

By Claim 2.5 we conclude that one can identify the types of the components V_1 and V_2. This means that we can identify at least $(1/2 - \delta)n$ truthful vertices with no error. Recall that this is the case also when H has only one connected component of size at least

$(1/2 - \delta)n$. Having these truthful vertices, we also know the types of all their neighbors. By the assumption on G this gives the types of all vertices but less than δn, completing the proof of the main part of the theorem.

It was easy to establish that the algorithm is linear provided $|T| > (1/2 + \delta)n$ is clear. However, if we only assume that $|T| > |B|$ the proof provides only a non-efficient algorithm for deciding the types of the components V_1 and V_2. Indeed, we have to compute the maximum weight of an independent set containing 1 in the weighted graph S, and the maximum weight of an independent set containing 2. By the proof above, exactly one of this maxima is smaller than $1/2$, providing the required types. □

We next show that the non-efficiency of the algorithm is necessary.

Proof of Theorem 1.2: The proof is based on the following fact [5]: there exist constants $b < a < 1/2$ such that deciding if a graph H on m vertices, all of whose degrees are bounded by 4, has a maximum independent set of size at least $(a + b)m$ or at most $(a - b)m$ is NP-hard.

Let G' be a δ good bounded degree expander on a set V of n vertices. Split the vertices into 3 disjoint sets V_1, V_2, V_3, where V_3 is an independent set in G' of size m, where $bm = \eta n$, all its neighbors are in V_2, $|V_1| = n/2 - am$ and $|V_2| = n/2 - m + am$. Add on V_3 a bounded degree graph H as above, in which it is hard to decide if the maximum independent set is of size at least $(a+b)m$ or at most $(a-b)m$. That is, identify the set of vertices of H with V_3 and add edges between the vertices of V_3 as in H. Call the resulting graph G and note that it is a δ-good expander (as so is its spanning subgraph G').

The reports of the vertices are as follows. Each vertex in V_1 reports true on each neighbor it has in V_1, and corrupt on any other neighbor. Similarly, each vertex of V_2 reports true on any neighbor it has in V_2 and corrupt on any other neighbor, and each vertex in V_3 reports corrupt on all its neighbors. Note that with these reports the connected components of the graph H in the proof of Theorem 1.1 are V_1, V_2 and every singleton in V_3.

It is easy to check that here if H has an independent set I of size at least $(a+b)m$, then G has a set T of truthful vertices of size at least $n/2 + bm$, namely, the set $I \cup V_1$, which is consistent with all reports. If H has no independent set of size bigger than $(a - b)m$, then G does not admit any set T of truthful vertices of size bigger than $n/2 - bm$ consistent with all reports. This completes the proof. □

Proof of Theorem 1.3: Let B' be a set of at most ϵn vertices of G whose removal splits G to connected components with vertex classes $V_1, V_2, \ldots, V_s$, each of size at most ϵn. Consider the following s possible scenarios R_i, for $1 \leq i \leq s$.

R_i: the set of corrupt vertices is $B = B' \cup V_i$, all the others are good vertices. The vertices in B' report that all their neighbors are corrupt. The vertices in V_i report that their neighbors in V_i are in T, and that all their other neighbors are in B. (The truthful vertices, of course, report truthfully about all their neighbors).

It is not difficult to check that in all these s scenarios, all vertices make exactly the same reports. On the other hand, there is no vertex of G that is truthful in all these scenarios, hence no deterministic algorithm can identify a truthful vertex with no error. Since the number of corrupt vertices in all scenarios is at most $2\epsilon n$, the first assertion of the theorem follows. The claim regarding planar graphs and graphs with excluded minors follows from the results in [10], [3]. □

2.2 Directed graphs

Proof of Lemma 1.4: The graphs constructed are orientations of undirected expanders. Here are the details. Let $G = (V, E)$ be a d-regular undirected non-bipartite Ramanujan graph, where $d = \Theta(1/\delta)$ (see [9]). This is a Cayley graph with $d/2$ generators, and its girth is bigger than $\frac{2}{3}\frac{\log n}{\log d}$. Let E_1 denote all edges corresponding to, say, $k = \lceil 3\sqrt{d} \rceil$ of the generators and their inverses. Thus (V, E_2) is a $2k$-regular graph, take an arbitrary Eulerian orientation of it (an orientation where each vertex has in-degree and out-degree k). Orient the rest of the edges randomly, that is, for each edge $e \in E - E_1$ choose, randomly, independently and uniformly, one of the two possible orientations. As shown in [2] the average degree in the induced subgraph of G on any set of γn vertices does not exceed $\gamma d + 2\sqrt{d-1} < 3\sqrt{d}$ provided $\gamma < 1/\sqrt{d}$. In particular, if $\gamma \leq 8\delta$ and $d < 1/\gamma^2$ (which holds in our case, as $d = \Theta(1/\delta)$), the above inequality holds. Now if U is any set of at most $\gamma n/2$ vertices, and $U' = N^+(U) - U$ satisfies $|U'| \leq |U|$, then the set $U \cup U'$ is of size at most γn but contains at least $k|U| \geq k(|U| + |U'|)/2$ edges: namely all the edges of E_1 emanating from some vertex of U. This means that the average degree in the induced subgraph on $U \cup U'$ is at least $k \geq 3\sqrt{d}$, which is impossible. This shows that our directed graph satisfies property (i) (independently of the orientation of the edges in $E - E_1$). To prove that (ii) holds with high probability note that for any fixed disjoint sets of vertices A and B of sizes $|A| \geq \delta n$ and $|B| \geq n/4$, the expander mixing lemma (c.f., e.g., [4], Corollary 9.2.5) implies that there are more than $2n$ edges of $E - E'$ connecting

A and B, provided d is at least some c/δ. The probability that all these edges are directed from A to B, or that all of them are directed from B to A is smaller than $2^{-(2n-1)}$. As the number of choices for the pair of sets A and B is much smaller than 2^{2n-1}, we conclude that our oriented graph satisfies (ii) as well with high probability, completing the proof of the lemma. □

Proof of Theorem 1.5: The proof resembles that of Theorem 1.1 but requires several additional ideas.

Let H be the spanning subgraph of G in which an edge (u, v) of G is an edge of H iff u reports that $v \in T$. Let $V_1, V_2, \ldots, V_s$ be the sets of vertices of the strongly connected components (SCCs, for short) of H.

Claim 2.6 *All the vertices of each V_i are of the same type, that is, for each $1 \leq i \leq s$, either $V_i \subset T$ or $V_i \subset B$.*

Proof: If $u \in T$ and v is an outneighbor of u in H, then $v \in T$ (as u reports so). If $v \in B$, and u is an in-neighbor of v in H, then $u \in B$ (as u reports that $u \in T$). □

Call an SCC of H truthful if it is a subset of T, else it is a subset of B and we call it corrupt.

Let H' be the induced subgraph of G on the set T of all truthful vertices.

Claim 2.7 *Any SCC of H' is also an SCC of H.*

Proof: If $u, v \in T$ and (u, v) is an edge of G, then it is an edge of H too. Thus each SCC C' of H' is contained in an SCC C of H. This SCC is truthful, by Claim 2.6, and cannot contain any additional truthful vertices as otherwise these belong to C' as well. □

Claim 2.8 *The graph H' contains an SCC of size at least $|T| - 2\delta n > (1/2 - 2\delta)n$.*

Proof: Consider the component graph of H': this is the directed graph F whose vertices are all the SCCs of H', where there is a directed edge from C to C' iff there is some edge of H' from some vertex of C to some vertex of C'. It is easy and well known that this graph is a directed acyclic graph, and hence there is a topological order of it, that is, a numbering $C_1, C_2, \ldots, C_r$ of the components so that all edges between different components are of the form (C_i, C_j) with $i < j$. Order the vertices of H' in a linear order according to this topological order, where the vertices of C_1 come first (in an arbitrary order), those of C_2 afterwards, etc. Let u_i be the vertex in place i according to this order ($1 \leq i \leq |T|$). If the vertices $u_{\delta n}$ and $u_{|T|-\delta n+1}$ belong to the same SCC, then this component is of size at least $|T| - 2\delta n$ and we are done. Otherwise, the SCC containing $u_{|T|/2}$ differs from either that containing $u_{\delta n}$ or from that containing $u_{|T|-\delta n+1}$. In the first case, the set A of all SCCs up to that containing $u_{\delta n}$ is of size at least δn, and the set B of all, SCCs starting from that containing $u_{|T|/2}$ is of size at least $|T|/2 \geq n/4$, and there is no edge directed from B to A, contradicting the property of G. The second case leads to a symmetric contradiction, establishing the claim. □

Note that the above shows that if $|T| > (1/2 + 2\delta)n$ then H' and hence also H must contain an SCC of size bigger than $n/2$, which must be truthful. Thus, if this is the case, more than $n/2$ of the truthful vertices of G can be identified by the known linear time algorithm that computes the strongly connected components of H ([15], see also [14]). In addition, since all vertices but less than δn are among their out-neighbors, this enables us to identify the types of all vertices besides less than δn).

We next show that even if we only assume that $|T| > n/2$ we can still identify correctly most of the truthful vertices.

By the last two claims it follows that H contains at least one SCC of size at least $(1/2 - 2\delta)n \geq 3/8n$. If H contains only one such component, then this component must consist of truthful agents, and we can identify all of them (and hence also the types of all their out-neighbors). Otherwise, there is another SCC of size at least $(1/2 - \delta)n$, and as there is no room for more than two such components, there are exactly two of them, say V_1 and V_2. Note that by the properties of G there are edges of G from V_1 to V_2 and from V_2 to V_1, and hence it is impossible that both of them are truthful components. As one of them must be truthful, it follows that exactly one of them is truthful and one is corrupt. We next show that we can identify the types of both components.

Recall that we have the SCCs of H, and the set T of all truthful vertices must be a union of a subset of these SCCs. In addition, this set must be of size bigger than $n/2$ and must be consistent with all reports of the vertices along every edge (in the sense that for any edge (u, v) with $u \in T$, the report of u on v should be consistent with the actual type of v.)

Claim 2.9 *Given the strongly connected components $V_1, V_2, \ldots, V_s$ of H and the reports along each edge, either there is no union I_1 of SCCs including V_1 whose size exceeds $n/2$ so that $T = I_1, B = V - I_1$ is consistent with all reports along the edges, or there is no union I_2 of SCCs including V_2 whose size exceeds $n/2$ so that $T = I_2, B = V - I_2$ is consistent with all reports along the edges.*

Proof: Assume this is false, and let I_1, I_2 be as above. By the above discussion we know that I_1 contains V_1 but not V_2 and I_2 contains V_2 but not V_1. Note that if some SCC V_i is contained both in I_1 and in I_2 and there is any directed edge (u, v) from V_i to some other SCC V_j, then if the report along this edge is that v is truthful, then V_j must be truthful component in both I_1 and in I_2. Similarly, if the report along this edge is $v \in B$, then V_j must be outside I_1 and outside I_2. In particular, there are no edges at all from V_i to V_1 or V_2 (as each of them lies in exactly one of the two unions I_1, I_2). Let J be the set of all SCCs that are contained in both I_1, I_2. By the remark above, for every edge (u, v) from a vertex of J to a vertex outside J, the report along the edge must be $v \in B$ (since otherwise v would also be in an SCC which is truthful both in I_1 and in I_2 and hence would be in J). Thus all edges (u, v) as above report $v \in B$, implying that all components outside J to which there are directed edges from vertices in J belong to neither I_1 nor I_2. By the properties of our graph the total size of these components exceeds that of J, (as $|J| \leq 4\delta n$ and all out-neighbors of J are outside V_1, V_2), and this shows that the sum of the sizes of I_1 and I_2 is at most

$$2|J| + (|V| - |J| - |N^+(J) - J|) \leq |V|.$$

Therefore it cannot be that both I_1 and I_2 are of size bigger than $n/2$, proving the claim. □

By the last claim it follows that one can identify the types of the SCCs V_1 and V_2. This means that we can identify at least $(1/2 - 2\delta)n$ truthful vertices with no error. Recall that this is the case also when H has only one SCC of size at least $(1/2 - \delta)n$. Having these truthful vertices, we also know the types of all their out-neighbors. By the assumption on G this gives the types of all vertices but less than δn, completing the proof of the main part of the theorem.

The comment about the linear algorithm provided $|T| > (1/2 + 2\delta)n$ is clear. If we only assume that $T| > |B|$ the proof provides only a non-efficient algorithm for deciding the types of the SCCs V_1 and V_2. Indeed, we have to check all 2^s possibilities of the types of each of the SCCs and see which ones are consistent with all reports and are of total size bigger than $n/2$. By the proof above, only one of the two SCCs V_1, V_2 will appear among the truthful SCCs of such a possibility. □

3 Discussion and Open Problems

Our results show that for very good expanders it is possible to find many of the truthful nodes even if there is only one more truthful than untruthful nodes. They also show that for graphs with very bad expansion properties (like an interval or a grid) even with a very high percentage of truthful nodes, it is hard to find even one truthful node. It is interesting to study in more detail the relation between expansion and corruption.

Question 3.1 *Provide sharp criteria in terms of expansion and the fractional size of the set T for corruption detection.*

To illustrate an example of such result, consider the following argument. We say that an undirected graph G is *δ-connected* if for every two disjoint sets A_1, A_2 with $|A_1| \geq \delta n, |A_2| \geq (1-3\delta)n$ there is at least one edge between A_1 and A_2. Note that the notion of connectedness is much weaker than expansion. In particular a graph G can be δ connected, yet at the same time have $\delta n/2$ isolated vertices.

Claim 3.2 *Suppose that $|T| = (1-\epsilon)n$ and the graph G is ϵ-connected then it is possible to identify $T' \subset T$ of size at least $(1-2\epsilon)n$.*

Proof: Let $E' \subset E$ be the set of edges both of whose end-points declare each other truthful. Recall that each connected components of $G' = (V, E')$ is either truthful or corrupt.

Let $T_1, T_2, \ldots$ denote all the components of size at least ϵn in G'. Then we claim that if $T' = \cup T_i$ then $|T \setminus T'| < \epsilon n$. Assume otherwise. Since all the connected components of $T \setminus T'$ are of size at most ϵn, there exists $T'' \subset T \setminus T'$ of size in $[\epsilon n, 2\epsilon n]$ with no edges to $T \setminus T''$ whose size is in $[(1-3\epsilon)n, (1-2\epsilon)n]$. This is a contradiction to ϵ-connectedness and the proof follows. □

To see that the conditions of Claim 3.2 are tight up to constant factors consider star graph with m leaves. Assume that $|T| \leq m-1$. Then it is easy to see that one cannot find even one member of T if all vertices declare all their neighbors corrupt. On the other hand, this example is (vacuously) $1/(4m)$ connected. To get a non-trivial example, one can replace each node with a complete graph K_k and each edge with a complete bipartite graph $K_{k,k}$ for an arbitrary k.

We conclude with a short discussion of a variant of the model. From the modeling perspective, it is interesting to consider probabilistic aspects of the corruption detection problem.

Question 3.3 *What is the effect of relaxing the assumption that truthful nodes always report the status correctly? Suppose for example that each truthful node reports the status of each of its neighbors independently accurately with probability $1-\epsilon$. Note that in this case it is impossible to detect the status of an individual node with probability one. However* *it is still desirable to find sets T' and B' such that the symmetric difference $T \Delta T'$ and $B \Delta B'$ are small with high probability. Under what conditions can this be achieved? What are good algorithms for finding T' and B'?*

Acknowledgment: We thank Gireeja Ranade for suggesting to consider problems of corruption on networks. We thank Peter Winkler for helpful comments regarding the machine puzzle and byzantine computing.

References

[1] N. Alon, J. Bruck, J. Naor, M. Naor and R. Roth, Construction of asymptotically good, low-rate error-correcting codes through pseudo-random graphs, IEEE Transactions on Information Theory 38 (1992), 509-516.

[2] N. Alon and F. R. K. Chung, Explicit construction of linear sized tolerant networks, Discrete Math. 72(1988), 15-19; (Proc. of the First Japan Conference on Graph Theory and Applications, Hakone, Japan, 1986.)

[3] N. Alon, P. D. Seymour and R. Thomas, A separator theorem for graphs with an excluded minor and its applications, Proc. 22^{nd} annual ACM Symp. on the Theory of Computing (STOC), Baltimore, Maryland, 1990, ACM Press, 293-299. Also: A separator theorem for non-planar graphs, J. Amer. Math. Soc. 3 (1990), 801-808.

[4] N. Alon and J. H. Spencer, *The Probabilistic Method, Third Edition*, Wiley, 2008, xv+352 pp.

[5] P. Berman and M. Karpinski, On some tighter inapproximability results, Proceedings of ICALP, LNCS 1644, Springer Berlin Heidelberg, 1999.

[6] O.H Fjeldstad, Fighting fiscal corruption: lessons from the Tanzania Revenue Authority, Public Administration and Development 23.2 (2003), 165-175.

[7] M. I. Steinder, and A. S. Sethi, A survey of fault localization techniques in computer networks, Science of computer programming 53.2 (2004), 165–194.

[8] L. Lamport, R. Shostak, and M. Pease. The Byzantine generals problem. ACM Transactions on Programming Languages and Systems (TOPLAS) 4.3 (1982): 382–401.

[9] A. Lubotzky, R. Phillips, and P. Sarnak, Ramanujan graphs, Combinatorica 8 (1988), 261–277.

[10] R. J. Lipton and R. E. Tarjan, A separator theorem for planar graphs, SIAM J. Appl. Math. 36 (1979), 177–189.

[11] B. Mukherjee, L. Todd Heberlein, and K. N. Levitt., Network intrusion detection, Network, IEEE 8.3 (1994), 26–41.

[12] R. P. Nielsen, Corruption networks and implications for ethical corruption reform, Journal of Business ethics 42.2 (2003), 125–149.

[13] M. T. Rock and H. Bonnett, The comparative politics of corruption: accounting for the East Asian paradox in empirical studies of corruption, growth and investment, World Development 32.6 (2004), 999–1017.

[14] M. Sharir, A strong-connectivity algorithm and its applications in data flow analysis, Comput. Math. Appl. 7 (1981), 67–72.

[15] R. E. Tarjan, Depth-first search and linear graph algorithms, SIAM Journal on Computing 1 (1972), 146–160.