Social Networks 54 (2018) 228–237

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Birds of a feather scam together: Trustworthiness homophily in a business network

Mauro Barone [a], Michele Coscia [b,c,*]

[a] Agenzia delle Entrate – Ufficio Studi Economico-Statistici, Via C. Colombo 426 c/d, 00145 Roma, Italy
[b] Naxys – University of Namur, Rempart de la Vierge 8, 5000 Namur, Belgium
[c] Center for International Development - Harvard University, 79 Jfk St, Cambridge 02138, United States

ARTICLE INFO

Article history:

Keywords:
Tax evasion
Fraud detection
Complex networks

ABSTRACT

Estimating the trustworthiness of a set of actors when all the available information is provided by the actors themselves is a hard problem. When two actors have conflicting reports about each other, how do we establish which of the two (if any) deserves our trust? In this paper, we model this scenario as a network problem: actors are nodes in a network and their reports about each other are the edges of the network. To estimate their trustworthiness levels, we develop an iterative framework which looks at all the reports about each connected actor pair to define its trustworthiness balance. We apply this framework to a customer/supplier business network. We show that our trustworthiness score is a significant predictor of the likelihood a business will pay a fine if audited. We show that the market network is characterized by homophily: businesses tend to connect to partners with similar trustworthiness degrees. This suggests that the topology of the network influences the behavior of the actors composing it, indicating that market regulatory efforts should take into account network theory to prevent further degeneration and failures.

1. Introduction

Suppose a judge has the task of conciliating two parties making different claims. If all the information available comes from the two parties, it is impossible to determine objectively where truth lies. However, if information about all cases regarding the two parties is public, it is possible to know which of the two is usually associated with larger mismatches – and likely to be less trustworthy.

In this paper, we show a simple formalization of this process using social networks. Each actor in the network is a source of reports about the other actors. Such reports constitute the edges of the network. The edges can contain mismatches: sometimes actor *a* reports something about its relationship with actor *b* that is not perfectly reciprocated. We develop an iterative framework to estimate the trustworthiness level of actors in a network when such mismatches are present.

We choose to focus on a real application scenario: the detection of tax fraud in a business-to-business (B2B) customer–supplier network. Each transaction running from a supplier to a customer carries a packet of information that can be used to estimate the degree of trustworthiness the business partners have. When mismatches arise because the partners disagree on the amount of their transaction, we have to solve the same ontological problem of our hypothetical judge: discerning the virtuous businesses from the fraudulent ones. We solve such problem by recursively updating the trustworthiness of a business with the trustworthiness of the partners with which it disagrees. The solution fits into the social network research branch dedicated to the estimation of node centrality in complex networks (Katz, 1953; Bonacich, 1987; Borgatti and Everett, 2006; Page et al., 1999), or to the detection of malicious bots in social media (Ferrara et al., 2016). In fact, our social network perspective allows for more than just identifying fraudulent nodes in the market system. We can investigate fundamental properties of the shadow market network. One such property is homophily: the actors in our network preferentially attach to actors with a comparable level of trustworthiness. In social systems, homophily is the tendency of actors to connect with other actors that are similar to them. Researchers have shown that this is a pervasive and ubiquitous aspect of social (McPherson et al., 2001; Mollica et al., 2003) and economic systems (Jackson, 2008), even virtual ones (Szell et al., 2010).

Note that our modeling is devoid of normative aspects: we do not advocate for a particular solution to the problem of fix-

* Corresponding author at: Naxys – University of Namur, Rempart de la Vierge 8, 5000 Namur, Belgium.
E-mail address: michele.coscia@hks.harvard.edu (M. Coscia).

https://doi.org/10.1016/j.socnet.2018.01.009
0378-8733/

ing tax fraud. However, the approach presented here can be seen as a building block of a theory that accounts for the process by which this phenomenon arises. During the last 50 years, models following classical and non-classical economic theory tried to understand how and why the shadow network of tax fraud arises (Allingham and Sandmo, 1972; Feige, 2007; Alm, 2012). Approaches to study the phenomenon range from game-simulation strategies (Friedland et al., 1978), to econometrics models based on behavioral hypotheses (Myles and Naylor, 1996), to fully-fledged behavioral economics models (Hashimzade et al., 2013; Granovetter, 2005). Our results show that, by extending these efforts with network theory – from the understanding of scale free effects (Barabási et al., 2000) to the detection of meso structures and functional modules (Rombach et al., 2014; Coscia et al., 2011) – we could paint a fuller picture of the informal sector and how to fix the resulting market inefficiency.

Our results are based on a network of 44,889 Italian businesses who reported their customers and suppliers in 2007. We examine several aspects of the spread of suspicious mismatches in these records. With an iterative mismatch correction algorithm, we quantify the degree of trustworthiness of each business, correcting biases in the baseline evaluation that are due to nodes in position of power in the network. We validate our measure of trustworthiness by showing that it is able to predict if a business is going to pay a fine for tax evasion if audited, and the amount of the fine itself. Finally, we show that there is an association between one business' trustworthiness score and the scores of its partners: an evidence that the market network is characterized by homophily.

2. Materials and methods

2.1. Data

Under Italian law, firms are required to record all business to business operations, regardless of the amount. This data is recorded in the customer and supplier lists, where each operation is connected to the partner business. The data is collected each year and used to check mismatches and deploy audits.

The *Agenzia delle Entrate* provided us the customer and supplier lists of a selected sample of businesses, focusing on the year 2007. We start from a seed list of 1559 audited subjects from a single Italian region (Tuscany). We then select all customers and suppliers of these 1559 businesses, ending up with a total node set of 44,887 subjects. We collect all business relations established among these nodes. The 43,328 businesses not part of the seed set have relations with subjects not included in the network, but for simplicity we consider our sample a closed system, since it contains all relations among the studied subjects. The assumption is that the external relations are on average no different than the sampled relationships.

To generate this initial dataset we had to solve issues about the same VAT numbers referring to different businesses identifiers, multiple reports provided by a business, and duplicated records. We detail our solutions in the Supplementary Material Section 1. Fig. 1 depicts a view of a full relationship between two hypothetical businesses (a and b). The set of all such relations composes the partnership network P. Note that, in this example, the two businesses agree about the amount b sold to a (75). However, they disagree on the amount a sold to b: b is under-reporting (95) and a is over-reporting (100). This disagreement is the basis of our analysis.

Table 1 reports basic statistics of the final dataset. Each pair is a business interaction between two businesses, where one business sold something – product or service – to another. The first business is the provider, the second is the customer. For each interaction, we have two reports: one from the point of view of the customer and one from the point of view of the provider. Note that the number of reports is lower than the double of the number of pairs: this means that there are some instances – ~7% of transactions – in which one of the two businesses failed to acknowledge the other party as a partner in a transaction. The transaction volume included in the dataset represents approximately 0.56% of Italy's GDP.

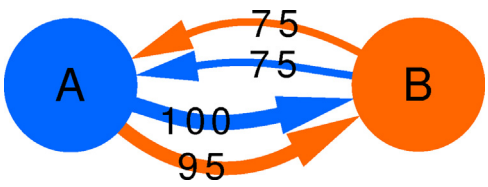

Fig. 1. The schema of our data structure for a full relationship. Businesses a and b are both suppliers and customers of each other. The direction of the edge goes from the supplier to the customer. The blue edges are reports from a, and the orange edges are reports from b. The amount reported is represented by the edge's label and thickness. So the orange edge from a to b is b's report about how much it bought from a, while the blue edge from a to b is a's report about how much it sold to b.

Table 1
The basic statistics of our dataset. We report the number of subjects (both in the seed set and in the total network); the number of expressed subject pairs (i.e. pairs of businesses that were suppliers, customers or both); the number of reports submitted, ideally two per pair (one from the customer and one from the provider); the total transaction volume in billions of Euro in the dataset, assuming the average of two conflicting reports is correct.

Variable	Value
Seed set size	1559
# Subjects	44,889
# Pairs	847,513
# Reports	1,578,121
Volume (Avg)	€9.094B

2.2. Trustworthiness

The principal task in this work is to establish the degree of trustworthiness of a business. There is a trivial solution to this problem: to calculate its average level of disagreement with all the businesses with which it interacts. We define the mismatch function for a pair of partnering businesses a and b as:

$$M(a,b) = |\alpha_a(a \to b) - \alpha_b(a \to b)| + |\alpha_a(b \to a) - \alpha_b(b \to a)|.$$

$\alpha_a(a \to b)$ denotes the value of the record reported by a of the amount sold by a to b. We define the operation volume of the pair as:

$$\Psi(a,b) = \alpha_a(a \to b) + \alpha_b(a \to b) + \alpha_a(b \to a) + \alpha_b(b \to a).$$

Now we can evaluate the ground trustworthiness function:

$$T_0(a,b) = 1 - \frac{M(a,b)}{\Psi(a,b)}.$$

$T_0(a,b)$ takes values between 0 and 1, where 1 means perfect agreement between a and b, and 0 means complete disagreement – either a or b did not acknowledge their partner. In the example from Fig. 1, $M(a,b) = 5$, $\Psi(a,b) = 345$, $T_0(a,b) \sim 0.9855$.

We can evaluate the overall trustworthiness of business a by calculating T_0 with respect to all its partners. We refer to this function as $T_0(a,\cdot)$, contracted as $T_0(a)$:

$$T_0(a) = \frac{1}{|N_P(a)|} \sum_{b \in N_P(a)} T_0(a,b),$$

where $N_P(a)$ is the set of all business partners (neighbors) of a in the partnership network P.

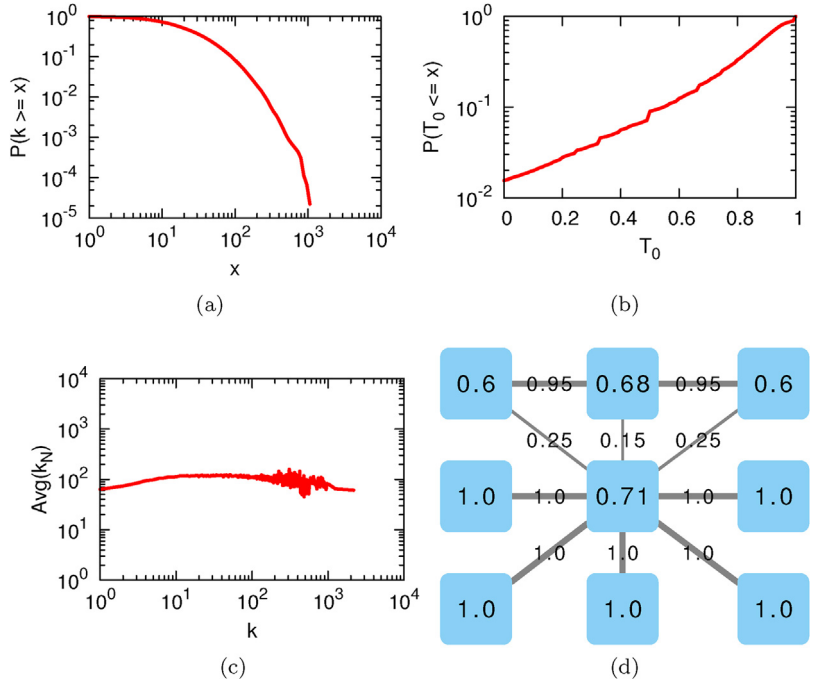

Fig. 2. Some topological properties of the P' network. (a) The cumulative degree distribution of P', where the x-axis reports the degree k – number of business partners, regardless if customers or suppliers – and the y-axis the probability for a node of having degree k or higher. (b) The distribution of trustworthiness T_0 for all edges in P'. The x-axis reports the value of T_0 and the y-axis the probability for an edge to have trustworthiness equal to or lower than T_0. (c) The relationship between a node's degree (x-axis) and its average neighbor degree (y-axis) in P'. Note that for all nodes with $k < 100$, which compose ~92% of the network, $k < Avg(k_N)$ – confirming the friendship paradox for P'. (d) An example of the hub avalanche effect. Each node is a business and each edge reports the agreement between the two reports. The node label is the node's T_0 score. We assume all edges have the same volume. Supposing that we know the cause of mismatches in the hub's connections is exclusively itself, the hub's T_0 score is still higher than its non-hubs victims, that are otherwise reliable. However, their few connections cannot counter balance the ones they have with the fraudulent hub.

For convenience, we define a projected view of P. For each pair of connected businesses, we collapse their multiple connections – up to 4 directed edges – into a single undirected edge. We assign to this edge, say (a, b), the trustworthiness score of the relationship: $T_0(a, b)$. We refer to this network as P'. For instance, while in P the businesses a and b from Fig. 1 have four edges of weight 75, 95, 100, and 75, in P' they have only one edge of weight 0.9855. In P', $T_0(a)$ can be interpreted as the average weight of a's connections.

2.3. Topology of the partnership network

Many real world networks are scale free, or have a generally broad degree distribution. This means that most of the nodes have a low number of connections, while there are large hubs connected to many nodes. This is true also for the partnership network P' connecting businesses with customer/supplier relationship. Fig. 2(a) depicts the degree distribution. P' is not scale free, but still more than 50% of its nodes have degree of 20 or less. The central hubs can have thousands of connections. Fig. 2(b) depicts how the baseline trustworthiness T_0 distributes: most businesses are honest, meaning that around 70% of businesses have a T_0 score higher than 0.8, i.e. they have a total match of 80%.

The two facts mean that, if we pick a node at random, we are likely going to pick a honest business with few business connections. According to the friendship paradox (Feld, 1991) – which holds in P', see Fig. 2(c) –, one of these businesses is likely to be a hub. Therefore, it takes a single fraudulent hub, however unlikely this is to happen, to propagate low T_0 scores unfairly to the small neighbors, while the hub itself might still have a high T_0 score, protected by the honest large quantity of small partners. In other words, a hub can shift the blame to its non-hub partners. Fig. 2(d) depicts an example of this avalanche hub effect.

This topological property also relates to an ontological issue of the simple T_0 measure. When we have a mismatch in the reports from a and b, without any source of external information we cannot know which business is to blame. T_0 distributes equally the blame to a and b even if, hypothetically, the source of the entire mismatch could be a.

2.4. Recursive trustworthiness

We address both issues by calculating the T_n score. T_n is an iterative correction of T_0. In practice, to evaluate the trustworthiness of a at step n, we make use of its trustworthiness – and the ones of all its partners – at step $n-1$. T_n uses T_0 as initial condition. It also has to respect the same constraint of T_0, namely to take values exclusively between 0 and 1.

To calculate $T_n(a, b)$, it is useful to first estimate the trustworthiness balance $B_n(a, b)$:

$$B_n(a, b) = \frac{T_{n-1}(a)}{T_{n-1}(b)}.$$

$B_n(a, b)$ is higher than 1 if, at the previous iteration, a was judged to be more trustworthy than b. We now plug the balance into the trust update:

$$T_n(a, b) = \begin{cases} \dfrac{B_n(a, b)}{B_n(a, b) + (1 - T_{n-1}(a))} & \text{if } T_{n-1}(b) \neq 0 \\ 1 & \text{if } T_{n-1}(b) = 0. \end{cases}$$

The two cases are required to deal with fully untrustworthy partners ($T_{n-1}(b) = 0$), as in that case $B_n(a, b)$ is undefined. Since b in this case was unreliable, we have no other choice than to trust a, hence $T_n(a, b) = 1$ and $T_n(b, a) = 0$.

Note that T_0 was symmetric ($T_0(a, b) = T_0(b, a)$), while $T_n(a, b)$ breaks this symmetry. This is because $B_n(a, b) \neq B_n(b, a)$, if

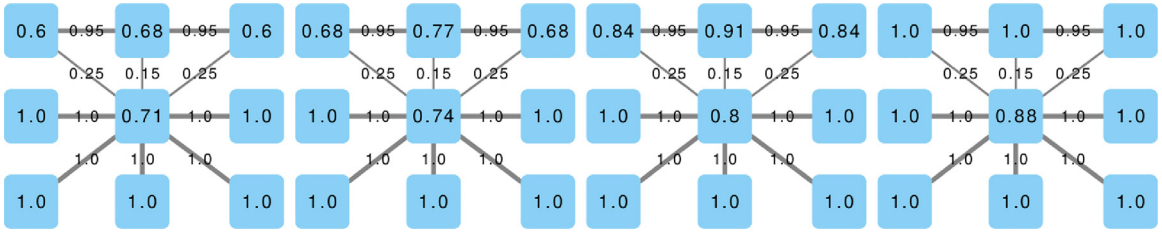

Fig. 3. The progression of T_n values of each node in the hub example for growing iteration indexes n. T_n values are used as node labels. From left to right: T_0 (starting condition), T_1 (first iteration), T_5, and T_{100}, which we chose as last iteration.

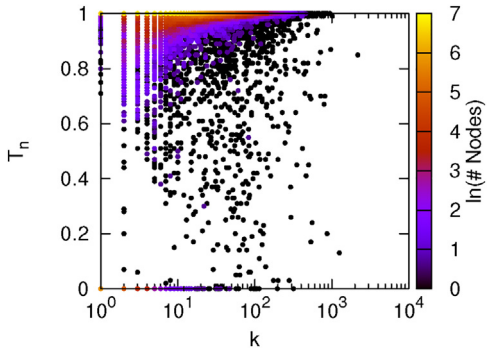

Fig. 4. Relationship between degree (x-axis) and T_n (y-axis). Each data point has been colored according to how many nodes (in logs) have the given degree and T_n values.

$T_{n-1}(a) \neq T_{n-1}(b)$ (just like $T_0(a)$, also $T_n(a)$ is the average of all $T_n(a, b)$). This is by design, as the asymmetry will transfer more trustworthiness to the more trustworthy side. $T_n(a, b)$ is proven to take values between 0 and 1, as any function of the form $\frac{x}{x+x_0}$ does, so long as $x_0 \geq 0$. In our case, this is true, as our x_0 is $1 - T_{n-1}(a)$. Since $T_0(a)$ takes values in between 0 and 1 by construction, $\nexists n$ such that $T_n(a)$ breaks the assumption.

Note that the overall trustworthiness in the system increases at each iteration. This is because, since $T_{n-1}(b) \leq 1$, then $B_n(a, b) \geq T_{n-1}(a)$. This is a desirable property. In practice, it means that if a business partner is not fully trustworthy then, however high the mismatch, the other party has to be acknowledged being partly in the right. However, if $T_{n-1}(b) = 1$, then a does not get any extra trustworthiness from the relation, and $T_n(a, b) = T_{n-1}(a, b)$.

Fig. 3 shows how T_n addresses the topological issue. In the first step of the iteration all scores increase, as expected since $T_n \geq T_{n-1}$. However, the speed at which the increase happens is slower for the central hub, given its not trustworthy nature. With $n = 5$, we have already a situation in which the central hub becomes the least trustworthy node in the network. When $n = 100$, there is virtually no more trust that can be exchanged in the system. The top nodes cannot have $T_n = 1$, but they are very close to it (>0.999). As a consequence, at each iteration the hub can increase its T_n by a negligible amount, and the system can be considered at convergence. The hub's privilege has been revoked.

T_n addresses the ontological issue by saying that, when evaluating the reports of a and b about each other, one should trust the business that has been judged the most trustworthy of the two so far. With its iterative updates, T_n tackles the recursive nature of this question.

2.5. T_n properties

T_n has a low correlation with the node's degree. Fig. 4 (left) shows that there is little to no relation between T_n and degree, as desired. In fact, the correlation between these two vectors is low (0.06). More importantly, the correlation is less than half the one between the degree and T_0, which is equal to 0.11. The small dependence of T_n with the degree is the reason why T_n is a better measure than PageRank, as the Spearman correlation of PageRank with degree is very high.

It is also important to assess how much T_n changes the estimates we had with T_0. If we update the values, but they maintain a strong correlation with T_0, then the operation is pointless. There is a significant linear correlation, equal to 0.63. This means that the two measures are related. However, this is expected and desired. On the other hand, T_n operates a significant reordering in the trustworthiness ranks, which is where its value resides. Top-trustworthy nodes can be demoted significantly and vice versa. This can be captured by estimating the Spearman rank correlation, which is significantly smaller than the linear correlation. It is equal to 0.34.

2.6. T_n convergence

Since we have not given proof of convergence for T_n, we need to show the asymptotic behavior of the function as $n \to \infty$. There are two measures of interest. First, the average difference between T_n and T_{n-1}. A well-behaved function would show an exponential decay as it approaches its final value. Second, the overall average T_n. Since we have shown T_n to be a monotone growing function, its final value might be $T_n = 1$ for all businesses. At this point, T_n would be useless.

We ran 100 iterations of T to estimate T_{100}. Fig. 5 depicts the result for both measures of interest. Both requirements are satisfied: the average difference across iteration does indeed decay exponentially, and the asymptote of T_n seems to lie below 0.98. Since there is little difference between iterations 99 and 100, we decide to fix $n = 100$ for the rest of the paper. Note that each iteration over more than 2 million records took an ordinary laptop a little more than 2 seconds with a fairly naïve Python implementation, demonstrating the time efficiency of the calculation. Each edge of P' is considered twice. That makes its time complexity $\mathcal{O}(ne)$, where e is the number of edges in P' and n is the number of iterations. If $e \gg n$, as it is the case if the network is large and convergence happens after few iterations, then the complexity is $\mathcal{O}(e)$. Being linear in term of the number of edges, the computation of T_n can be applied to very large networks with hundreds of millions of connections.

Fig. 6 depicts the evolution in the distribution of T_n as n grows. More nodes obtain higher T_n scores as they tend to be classified as trustworthy. However, the speed at which the function tends to its asymptotes slows down until it converges to its final form. All nodes for which $T_n < 0.99$ can be considered not trustworthy.

3. Results

3.1. Avalanche effect in the data

In the previous section we provided a topological argument for the hub avalanche effect. Is this effect actually affecting a signifi-

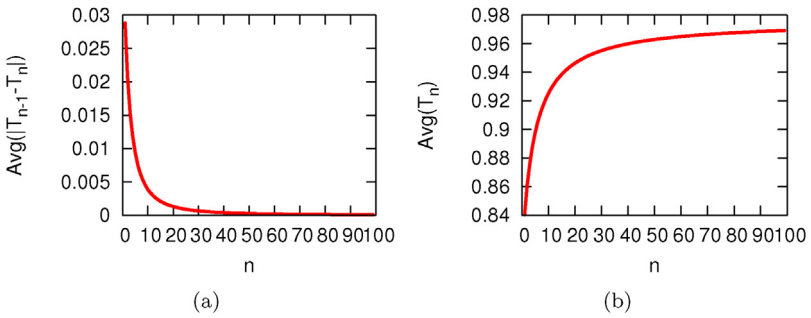

Fig. 5. Convergence of T_n: the asymptotic behavior for two measures of interest across iterations of T_n. (a) For each iteration n (x-axis) we report the average inter-iteration difference ($|T_n - T_{n-1}|$) on the y-axis. (b) For each iteration n (x-axis) we report the average T_n value of the businesses in the dataset (y-axis).

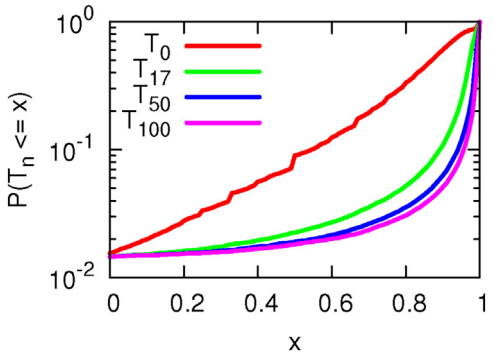

Fig. 6. Distributions of T_n at various n. Just like in Fig. 2(b), we report the probability (y-axis) of a business to have a T_n value (x-axis), for different ns.

cant portion of nodes in our data? To answer this question let us consider a hub as being a node with more than k connections. We then consider the neighbors of each hub. We define a victim of the avalanche effect as a business with lower T_0 than the hub, and either more than 90% of its other connections have a score higher than the hub's T_0, or the victim has no other connection.

We check multiple values of k (from 100 to 600). For each threshold, the share of victims among hub's neighbors varies between 12% and 16%. This means that there are thousands of small businesses in the network affected by the avalanche effect of untrustworthy hubs.

3.2. Homophily

Homophily is defined as the tendency of individuals to associate and bond with similar others (McPherson et al., 2001). Narrowing to a network perspective, homophily – also known as assortativity – implies that nodes will likely connect if they share similar characteristics. In our scenario, homophily implies that businesses with similar T_n scores are more likely to connect. If a business is honest it will tend to have honest customers/suppliers, and the converse applies if the business is not honest. We start by showing that T_0 implies homophily. Then we show that T_n – designed to abstract P' from assortative mixing – still shows all signs of homophily. This suggests that assortativity is an intrinsic characteristic of P'.

It is easy to see that T_0 implies homophily by construction. Since any mismatch between a and b will bring down both $T_0(a)$ and $T_0(b)$ by equal amounts, the T_0 estimation of two nodes at the endpoints of an edge is highly correlated. We provide three arguments in favor of this statement.

First, we focus on the difference of T_0 scores between businesses (ΔT_0). We compare the average ΔT_0 for connected pairs with the average ΔT_0 of unconnected pairs. Since there are too many unconnected pairs, we select a random sample of them of equal size of the number of connected pairs (~1.1 millions). We repeat the analysis 20 times and we report the average result. The way we select random unconnected pairs is the following: for each node in the network we pick a random set of non-neighbors of roughly equal size of the node's degree. The average ΔT_0 for neighbors is 0.1634. The average ΔT_0 for non-neighbors is 0.1976 (with standard deviation 2.35×10^{-4}). Since the two averages are significantly different, we use this as first proof of T_0 homophily in the network.

Table 2
The results of the model showing T_0 homophily in P'.

	Dependent variable:		
	ϵ		
	(1)	(2)	(3)
ΔT_0	−1.481***		−1.106***
	(0.008)		(0.009)
$\Delta\Psi$		0.396***	0.388***
		(0.001)	(0.001)
Region		0.964***	0.938***
		(0.004)	(0.004)
Industry		1.618***	1.571***
		(0.015)	(0.016)
Constant	0.248***	−5.603***	−5.306***
	(0.002)	(0.013)	(0.013)
Observations	1,694,297	1,694,297	1,694,297
Log Likelihood	−1,157,201.000	−1,032,081.000	−1,024,088.000
Akaike Inf. Crit.	2,314,407.000	2,064,169.000	2,048,185.000

Note:
*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01$.

As second argument, we show in Fig. 7(a) the relationship between ΔT_0 for neighbors and non-neighbors. The depicted ΔT_0 for non-neighbors is the average one we got over our 20 iterations. Only 12,552 businesses are below the identity line, while 32,337 lie above. This implies that for more nodes the difference in T_0 with neighbors is lower than with non-neighbors, an evidence for homophily.

In the final argument, we create a logit model with the aim of predicting the presence/absence of an edge in the partnership network P' by using ΔT_0. Table 2 reports the result of this model. The target variable is ϵ:

$$\epsilon(a,b) = \begin{cases} 0 & \text{if } (a,b) \notin P' \\ 1 & \text{if } (a,b) \in P'. \end{cases}$$

Again, we create a table where we introduce a set of random non-edges, of approximately equal size of the set of edges. When we test the relationship between ϵ and ΔT_0, we obtain a negative and significant relationship: the higher ΔT_0 the less likely the two businesses connect (model 1 in Table 2).

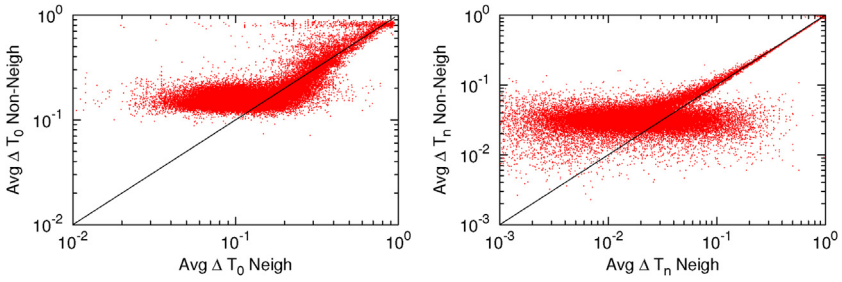

Fig. 7. The homophily plots for T_0 and T_n. (a) Homophily plot for T_0. Each data point is a business, On the x-axis we report its average T_0 difference with all its neighbors in P'. On the y-axis, its average T_0 difference with non-neighbors, averaged over 20 random extractions. The black line is the identity function $f(x) = x$. Homophily would imply that more points lie above the identity line than below, because their ΔT_0 with non-neighbors (y-axis) is higher than the ΔT_0 with neighbors (x-axis). (b) Same plot, but using T_n instead of T_0.

We have to control for other sources of homophily in the network: businesses do not connect with each other randomly. They are more likely to be each others customer/supplier if they are located in the same region, if they operate in the same industry, and all businesses – no matter how small – have to connect to very large providers of fundamental services (electric energy, telecommunication, etc.) implying a positive correlation with difference in volume – or $\Delta\Psi$. In fact, all these relationships are present and significant – as Table 2 model 2 shows. Table 2 model 3 shows that the T_0 homophily argument survives even when controlling for all these factors.

To gauge the substantive significance of this result, consider that the baseline probability of two nodes to be connected is 50% (since we focus on a balanced dataset containing as many random non-edges as edges). If ΔT_0 is in the bottom quartile (<0.04) the connection probability is 64%, while with $\Delta T_0 > 0.2$ (top quartile) this probability drops to 37%, suggesting a strong effect.

We say that T_0's assortativity is a weak indicator of homophily because it is derived directly by the way T_0 is computed. If we still find assortativity with T_n, then we can call it a strong evidence of homophily, because T_n actively fights assortativity. Take Fig. 3 as an example. In the starting condition on the left, there is an equal amount of edges between alike T_0 nodes than nodes with different scores. At the 100th iteration, almost all edges are between nodes with different T_{100} score. The degree of homophily in the simplified example is reduced, at the point of it being disassortative, rather than assortative. We then expect that, in a network without structural homophily, T_n will show signs of disassortativity (this is supported by simulation results reported in Section 3.6).

However, when substituting ΔT_0 with ΔT_n, all the three arguments we made previously for T_0's homophily still hold. The average ΔT_n for neighbors is 0.0491. The average ΔT_n for non-neighbors equals to 0.0552 (with standard deviation 2.21×10^{-4}, over the usual 20 iterations). Again, we see a significant difference, even if the scores are closer than before, because T_n values are skewed towards 1.

Fig. 7(b) shows the scatter of the average neighbor and non-neighbor ΔT_n scores. Again, we have more points above the identity line than below. The difference here is actually higher than before: 11,070–33,819 versus T_0's 12,552–32,337. The businesses tend to be closer to the identity line, but more of them end up above it.

Table 3 reports for ΔT_n the three models we calculated for ΔT_0 in Table 2. Again, model 3 shows a significant negative coefficient for ΔT_n, even controlling for the other sources of assortativity-disassortativity in P'. The effect magnitude is reduced (bottom ΔT_n quintile connection probability is 52%, top quintile is 40%), which is understandable given the different nature of T_n and T_0 when it comes to homophily, but it is still non-trivial. From these three arguments we can conclude that, no matter if we use T_0 or T_n, trustworthiness homophily is an intrinsic characteristic of P'.

Table 3
The results of the model showing T_n homophily in P'.

	Dependent variable:		
	ϵ		
	(1)	(2)	(3)
ΔT_n	−1.039***		−0.574***
	(0.011)		(0.012)
$\Delta\Psi$		0.396***	0.393***
		(0.001)	(0.001)
Region		0.964***	0.953***
		(0.004)	(0.004)
Industry		1.618***	1.611***
		(0.015)	(0.015)
Constant	0.045***	−5.603***	−5.541***
	(0.002)	(0.013)	(0.013)
Observations	1,694,297	1,694,297	1,694,297
Log Likelihood	−1,169,663.000	−1,032,081.000	−1,030,942.000
Akaike Inf. Crit.	2,339,331.000	2,064,169.000	2,061,894.000

Note:
*$p < 0.1$.
**$p < 0.05$.
*** $p < 0.01$.

3.3. Prediction quality

We now turn our attention to the practical application of T_n. We focus on the task of estimating the likelihood that a given business is maliciously misreporting their activities, evading their fiscal duties. An instrument to improve such task is beneficial for all actors in society. The *Agenzia delle Entrate* can perform audits with higher confidence of success, the government will be more efficient in upholding its fiscal laws, and the non-malicious businesses are less likely to lose time and resources to deal with an audit that should not have happened.

We divide the task in two parts. The first part focuses on predicting the probability that a business will have to pay a fine if checked. The second part aims to estimate how much the business will have to pay in fines, once it has been established as malicious. We perform the first part in this section and the second part in Section 3.4.

For both prediction tasks, we have to narrow down onto the initial seed set from Tuscany, composed by 1559 businesses, about which the data providers shared the audit results. All 1559 businesses were audited in 2007 and 1233 of them had to pay a fine. The distribution of fines is skewed, and Fig. 8(a) depicts it.

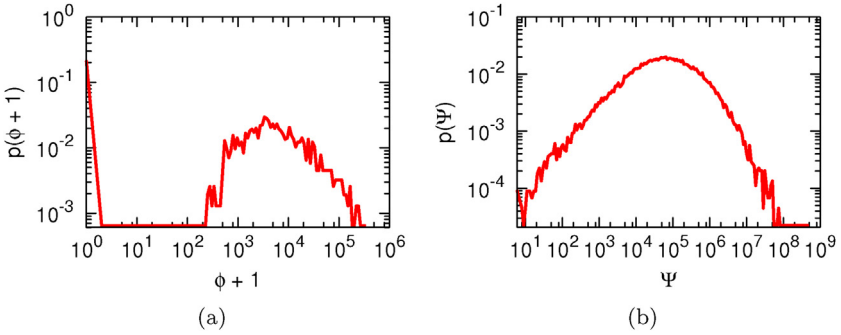

Fig. 8. The distributions of amount of fines (ϕ) and business volume (Ψ). (a) The probability distribution of having a given fine ϕ per business, among the 1559 seed set. Note that most businesses did not get a fine, so $\phi = 0$. To show the skewed distribution of ϕ, we increase ϕ by one and plot it in a log–log space. The large gap between 0 and 200 means that the *Agenzia delle Entrate* gave a negligible amount of fines lower than €200. (b) The probability distribution of having a given Ψ per business, spanning several orders of magnitude.

Table 4
The predictive power of T_n in estimating the probability that a business will be fined, if checked (logit model).

	Dependent variable:		
	$\bar{\phi}$		
	(1)	(2)	(3)
Ψ	0.116***	0.190***	0.171***
	(0.037)	(0.040)	(0.041)
T_0	0.0004		0.007*
	(0.003)		(0.004)
T_n		−0.039***	−0.047***
		(0.012)	(0.014)
Constant	0.091	3.125***	3.585***
	(0.366)	(1.163)	(1.261)
Observations	1559	1559	1559
Log Likelihood	−793.387	−785.139	−783.381
Akaike Inf. Crit.	1592.773	1576.279	1574.763

Note:
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

For the first task, we set up a logit regression. As dependent variable, we generate a binary variable from the amount of the fine. That is, suppose $\phi(a)$ is the amount a business had to pay in fine, the dependent variable $\bar{\phi}(a)$ is defined as:

$$\bar{\phi}(a) = \begin{cases} 0 & \text{if } \phi(a) = 0 \\ 1 & \text{if } \phi(a) > 0. \end{cases}$$

We test the effect of both T_0 and T_n in isolation in models (1) and (2), and then combine them in our full model (3):

$$\text{logit}\{P(\bar{\phi}(a) = 1 | X = x)\} = \alpha + \beta_1 \log(\Psi(a)) + \beta_2 T_0(a) + \beta_3 T_n(a) + \epsilon_a,$$

where α is the constant intercept and ϵ the error term. Note that in all models we control for the volume of business a ($\Psi(a)$). We expect the size of a business to play an important role in determining whether a business will misreport some of its transactions. Large businesses are more likely to do so even if by pure chance: the larger the volume of records, the more likely there will be errors in the data. This means that Ψ is a possibly significant confounding factor for which we need to account. The distribution of total business volume in the network spans several orders of magnitude. In fact, Ψ's value distribution is both left and right skewed. Fig. 8(b) depicts it. To account for this, in our model we take its natural logarithm.

Table 4 reports the results of models 1–3. Before interpreting the coefficients, we have to note that we are unaware of the current strategy used by the *Agenzia delle Entrate* to detect the malicious businesses in the network. For this reason, we are unable to correct possible confounding factors they might introduce in our results. It is likely that *Agenzia delle Entrate* is applying a mismatch strategy similar to T_0, since a logit model predicting the odds of being audited has a significant negative coefficient for T_0 (note that T_0 and T_n represent the *trustworthiness* of a business: the lower the T_0 trustworthiness of a business the more likely the business is going to be audited). If that is correct, we should not read too much in T_0's significance level and effect size, because we are already looking at businesses selected using the variable itself.

T_n's coefficient is negative and significant: the lower the T_n trustworthiness of a business the more likely that business is to be fined, if audited. The conclusion is that, by estimating T_n, we are introducing relevant information that can help discern better between trustworthy and non trustworthy businesses. In Section 3.5 we show through simulations that T_n is a good predictor of actual trustworthiness when removing the sample bias problem.

When interpreting the β coefficients note that we multiplied both T_0 and T_n to 100 for easier interpretability: the coefficients represent the change in odds for each percentage point increase in either T_0 and T_n. Focusing on model 2, for each percentage point increase in T_n the odds of being fined go down by $e^{-0.039}$. The baseline probability of being fined – controlling for size – if audited is ∼75.75 % (3.125/(3.125 + 1)), this means that having a T_n 10 percentage points lower than the average implies a fining probability of ∼77.85%. Using T_0 and T_n together (model 3) increases T_n predicting power, by controlling for the initial condition of honest businesses linked to malicious ones (T_0). In the same scenario presented before, the probability of being fined if audited with a T_n 10 percentage points lower than average would be ∼78.23%.

This is confirmed by empirical analysis. A business in the bottom T_n quartile has a probability of being fined if audited of ∼82.6%. A business in the top trustworthiness quartile has a lower fine probability: only 72.2%.

Fig. 9 reports the log likelihood of the model for different choices in the iteration parameter n. The quality of the model peaks for $n = 17$, which means that this is the value for which T_n is the most orthogonal with T_0 and therefore yields the best prediction. However, we are interested in the system at convergence, not in maximizing the prediction quality, and that is why we fix $n = 100$.

3.4. Predicting fine amount

After proving that T_n is a good predictor of the probability of paying a fine if audited, we now restrict our view only on those businesses that were audited *and* fined. The objective is to test whether T_0 and/or T_n are also significant predictors of the amount of the

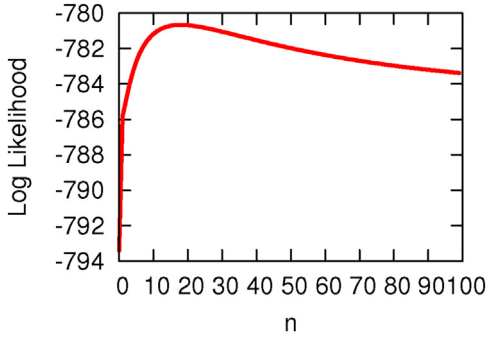

Fig. 9. The likelihood of the model predicting informality as a function of n in T_n (the number of iterations).

Table 5
The predictive power of T_n in predicting the amount a business will be fined, if found guilty of tax fraud.

	Dependent variable:		
	ϕ		
	(1)	(2)	(3)
Ψ	0.213***	0.228***	0.250***
	(0.026)	(0.027)	(0.027)
T_0	−0.015***		−0.010***
	(0.002)		(0.003)
T_n		−0.032***	−0.023***
		(0.005)	(0.006)
Constant	7.299***	9.183***	8.829***
	(0.262)	(0.446)	(0.453)
Observations	1233	1233	1233
R^2	0.059	0.062	0.072
Adjusted R^2	0.058	0.060	0.070
Residual Std. Error	1.507	1.505	1.497
F-statistic	38.833***	40.335***	31.890***

Note:
*$p < 0.1$.
**$p < 0.05$.
*** $p < 0.01$.

fine. We test a series of models similar to the ones in the previous section, testing T_0 and T_n first separately and then jointly:

$$\log(\phi(a)) = \alpha + \beta_1 \log(\Psi(a)) + \beta_2 T_0(a) + \beta_3 T_n(a) + \epsilon_a.$$

Also in this case, we control for the business' volume $\Psi(a)$. The rationale is that a large business is likely to pay a larger fine, since the *Agenzia delle Entrate* will take into account the business' size when deciding the amount of the fine. This also implies that the fines have a skewed distribution (Fig. 8(a)), and that is our reason for log-transforming the dependent variable ϕ.

Table 5 reports the results of these models. In this case, there is no disagreement between T_0 and T_n, as they both carry negative sign. The more trustworthy a business was, the less it is going to be fined if audited and found guilty. The effect size is small, as witnessed both by the coefficients and by the low R^2, although it is not a difference to scoff at: a bottom trustworthiness quartile fine averages at ∼17k€, while a top trustworthiness quartile fine averages at ∼13.7k€, a difference of more than three thousand Euros per audit. If we substitute 300 high trustworthiness audits with 300 low ones, the increased revenue would scratch a million Euros, even ignoring the fact that low T_n businesses also have a lower baseline probability to come out clean, as shown in the previous section. Moreover, the contribution of T_n is significant even in model 3. We conclude that T_n is a useful variable also for the task of predicting fine amounts, as it complements volume (Ψ) and simple network mismatch (T_0).

Table 6
Results of the simulation correlating the T scores with the hidden real $\mathcal{T}$ value.

	Dependent variable:		
	$\mathcal{T}$		
	(1)	(2)	(3)
Ψ	−0.008***	−0.023***	−0.013***
	(0.001)	(0.0005)	(0.0005)
T_0	0.531***		0.282***
	(0.003)		(0.004)
T_n		0.867***	0.620***
		(0.004)	(0.005)
Constant	0.285***	−0.150***	−0.117***
	(0.004)	(0.005)	(0.005)
Observations	100,000	100,000	100,000
R^2	0.247	0.296	0.336
Adjusted R^2	0.247	0.296	0.336
Residual Std. Error	0.251	0.242	0.235
F-statistic	16,363.350***	20,991.230***	16,853.850***

Note:
*$p < 0.1$.
**$p < 0.05$.
*** $p < 0.01$.

3.5. Simulation: prediction quality

One of the downsides of real data is that we can know whether a business is fined or not only for those who are audited. The *Agenzia delle Entrate* cannot audit all businesses, so it has to sample the ones that are most likely to pay a fine. Our sample is biased. We need to test how well our measure performs without a biased sample, but we cannot use real data given the constraint just stated.

One way to test T_n is by running a simulation. We generate a network with similar topological characteristics of P, we assign a secret real trustworthiness score $\mathcal{T}$ to each node at random, we modify each node record according to its trustworthiness and we apply our framework. If T_n can recover the information on $\mathcal{T}$ looking at all nodes in the network then it is a good measure.

As for the first step, we create a scale free directed graph. We fix the number of nodes at 1000. The graph is directed and weighted: each edge weight represents the volume of the transaction. The edge weights are chosen uniformly at random, between 1 and 100.

Then, we assign a secret trust score to each node of the network. Also these scores are extracted uniformly at random between 0 and 1, included. These are the real trust scores, or $\mathcal{T}$. We iterate over all edges in the graph and we generate a report by the two nodes involved in the relationship, a and b. Both a and b with a random 50% probability will either over-report or under-report. If a is under-reporting, we multiply the actual edge weight by $\mathcal{T}(a)$. If a is over-reporting, we multiply the actual edge weight by $1/\mathcal{T}(a)$. We perform the same operation for b, so we will have a mismatch, except in special cases – if $\mathcal{T}(a) = \mathcal{T}(b)$ AND they both under- or over-report, or $\mathcal{T}(a) = \mathcal{T}(b) = 1$ (both rather extreme cases). The result is a data structure equivalent to P.

We then generate P' from P, and calculate T_0 and T_n. We repeat the process 100 times, to account for random fluctuations. In Table 6 we report the results of the same regression we ran to estimate ϕ, but this time the dependent variable is $\mathcal{T}$, the real trust score that has not been used to build T_n. We also consider all nodes in the network, without sampling.

We can see that T_n is a good predictor of $\mathcal{T}$, confirming that it is indeed recovering the real trustworthiness values. The average $\mathcal{T}$ in the bottom quartile of T_n is just 0.2, while it is 0.61 in the top T_n quartile. In this case, also T_0 is positive and significant. This comes as no surprise, because we argue that T_0 is not a bad predictor, but the one the *Agenzia delle Entrate* is using to deploy the audits, thus biasing the sample in such a way to lower its coefficient below significance. Note that, when controlling for T_0, the effect of T_n is

Table 7
Results of the logit regression distinguishing edges and non-edges using the trustworthiness scores.

	Dependent variable:		
	ϵ		
	(1)	(2)	(3)
$\Delta\Psi$	−0.001	0.016***	−0.004***
	(0.001)	(0.001)	(0.001)
$\Delta\mathcal{T}$	0.0001		
	(0.006)		
ΔT_0		−0.510***	
		(0.011)	
ΔT_n			0.146***
			(0.014)
Constant	0.006	−0.038***	0.029***
	(0.007)	(0.007)	(0.007)
Observations	1,988,828	1,988,828	1,988,828
Log Likelihood	−1,378,550.000	−1,377,458.000	−1,378,496.000
Akaike Inf. Crit.	2,757,106.000	2,754,923.000	2,756,998.000

Note:
*$p < 0.1$.
**$p < 0.05$.
*** $p < 0.01$.

lower, because of the small co-linearity between the two scores. This seems to suggest that the T_n we observe in the prediction task using the real data might be an underestimation: in an unbiased sample T_n's coefficient should be higher.

3.6. Simulation: homophily

We perform a similar validation also for the homophily argument. We argue that T_0 is by design going to find assortativity in a non-assortative network, while T_n actively fights homophily and will, if anything, find disassortativity in a non-assortative network.

We generate 100 Erdos-Renyi random directed graphs, with $n = 1000$ and $p = 0.01$. Again, the real $\mathcal{T}$ scores are extracted uniformly at random. We apply the same procedure described above to derive P, P', T_0 and T_n. Then, we generate a set of non-edges of equal size of the edge set. We run a logit regression trying to predict which one is an actual edge using $\Delta\mathcal{T}$, ΔT_0 and ΔT_n. Since the network is completely random and the $\mathcal{T}$ scores are random too, the network is by definition non-assortative.

Table 7 reports the results. Consistently with the non-assortative nature of the network, $\Delta\mathcal{T}$'s coefficient is zero. ΔT_0's coefficient is negative and significant. This provides evidence in favor of our argument: we indeed find homophily in T_0's distribution even in a network that is non-assortative by definition. This is derived by how T_0 is calculated: whenever there is a mismatch, connected nodes will get an equal penalty in their T_0 scores.

However, ΔT_n's coefficient is positive and significant. The size of the effect is not of substantive significance, but statistical significance alone here is enough to sustain our argument: T_n by construction is trying to remove as much as possible homophily from the network. It goes so far in this attempt, that a non-assortative network appears to be disassortative, the opposite of homophily. The fact that we still find homophily in the business network is therefore even more remarkable. We leave the investigation on the origin of this strong homophily as future work.

4. Discussion

In this paper we develop a complex system framework to study the dynamics of tax fraud in a network of business partnerships. We are inspired by the applications of network analysis on evaluating the systemic risk of the global economy (Schweitzer et al., 2009), and on attempting to explain the causes of economic growth and predict it (Hausmann et al., 2011; Hidalgo and Hausmann, 2009). Network analysis has also been applied at the micro level to explain the retail behavior of single customers (Pennacchioli et al., 2014). Here, we apply this micro perspective to the task of evaluating for each business its likelihood of engaging in fraudulent transactions and tax evasion.

Tax evasion is a topic that has been studied extensively in the past (Allingham and Sandmo, 1972; Feige, 2007). A good recent review can be found in Alm (2012). Approaches to study the phenomenon range from game-simulation strategies (Friedland et al., 1978), to econometrics models based on behavioral hypotheses (Myles and Naylor, 1996), to fully-fledged behavioral economics models (Hashimzade et al., 2013). Most models, and especially the latter example, agree on the importance of the social element in the decision of evading taxes. However, none of them had the possibility of analyzing a dataset as detailed and as large as the one we present in this paper. None of them also did it using an approach rooted in network and complexity science, which we prove to be useful. Our contribution also provides a micro-level predictive system instead of modeling the overall distribution of tax evading behavior.

We provide this contribution by using mismatches in the network edges to estimate the degree of trustworthiness of businesses. We show that the trivial solution of averaging the mismatches has several problems. The main two are: the topology of the network implies an unbalanced estimation, giving an unfair advantage to high-degree hubs; and averaging mismatches is unable to detect the actual malicious actor, distributing the blame equally to all parties involved. We correct the mismatch average iteratively, creating a new trustworthiness score that is able to arbitrage trust in case of mismatched reports by looking at the network as a whole. This new score has a significant predictive power on the likelihood of being fined if audited. It also unveils fundamental properties of trustworthiness in a business network: businesses tend to connect with similar businesses. In other words, tax fraud homophily is a fundamental characteristic of business partnerships.

Acknowledgements

The authors thank the Agenzia delle Entrate for providing the data. We acknowledge Diego Pennacchioli for his role in the initial phases of this project, aiding in designing the experiments and performing preliminary analyses. The authors thank Renaud Lambiotte, Andres Gomez, Clara Vandeweerdt, Alan Szepieniec, Frank Neffke and Sebastian Bustos for useful discussions. M.C. has been partly supported by FNRS, grant #24927961.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.socnet.2018.01.009.

References

Allingham, M.G., Sandmo, A., 1972. Income Tax Evasion: A Theoretical Analysis.
Alm, J., 2012. Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. Int. Tax Public Financ. 19 (1), 54–77.
Barabási, A.-L., Albert, R., Jeong, H., 2000. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A 281 (1), 69–77.
Bonacich, P., 1987. Power and centrality: a family of measures. Am. J. Sociol. 92 (5), 1170–1182.
Borgatti, S.P., Everett, M.G., 2006. A graph-theoretic perspective on centrality. Soc. Netw. 28 (4), 466–484.
Coscia, M., Giannotti, F., Pedreschi, D., 2011. A classification for community discovery methods in complex networks. Stat. Anal. Data Min. 4 (5), 512–546.
Feige, E.L., 2007. The Underground Economies: Tax Evasion and Information Distortion. Cambridge University Press.

Feld, S.L., 1991. Why your friends have more friends than you do. Am. J. Sociol. 146, 4–1477.
Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A., 2016. The rise of social bots. Commun. ACM 59 (7).
Friedland, N., Maital, S., Rutenberg, A., 1978. A simulation study of income tax evasion. J. Public Econ. 10 (1), 107–116.
Granovetter, M., 2005. The impact of social structure on economic outcomes. J. Econ. Perspect. 19 (1), 33–50.
Hashimzade, N., Myles, G.D., Tran-Nam, B., 2013. Applications of behavioural economics to tax evasion. J. Econ. Surv. 27 (5), 941–977.
Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., Yildirim, M.A., 2011. The Atlas of Economic Complexity: Mapping Paths to Prosperity.
Hidalgo, C.A., Hausmann, R., 2009. The building blocks of economic complexity. Proc. Natl. Acad. Sci. U.S.A. 106 (26), 10570–10575.
Jackson, M.O., 2008. Social and Economic Networks. Princeton University Press.
Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18 (1), 39–43.
McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 41, 5–444.
Mollica, K.A., Gray, B., Trevino, L.K., 2003. Racial homophily and its persistence in newcomers' social networks. Organ. Sci. 14 (2), 123–136.
Myles, G.D., Naylor, R.A., 1996. A model of tax evasion with group conformity and social customs. Eur. J. Polit. Econ. 12 (1), 49–66.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The Pagerank Citation Ranking: Bringing Order to the Web.
Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F., Pedreschi, D., 2014. The retail market as a complex system. EPJ Data Sci. 3 (1), 1–27.
Rombach, M.P., Porter, M.A., Fowler, J.H., Mucha, P.J., 2014. Core-periphery structure in networks. SIAM J. Appl. Math. 74 (1), 167–190.
Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., White, D.R., 2009. Economic networks: the new challenges. Science 325 (5939), 422.
Szell, M., Lambiotte, R., Thurner, S., 2010. Multirelational organization of large-scale social networks in an online world. PNAS 107 (31), 13636–13641.