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.

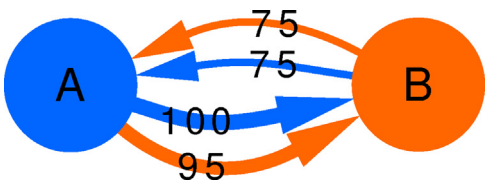

Fig. 1. The schema of our data structure for a full relationship. Businesses a and b are both suppliers and customers of each other. The direction of the edge goes from the supplier to the customer. The blue edges are reports from a, and the orange edges are reports from b. The amount reported is represented 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.