RECSM
Research and Expertise Centre for Survey Methodology

Implementing Propensity Score Matching with Network Data: The effect of GATT on bilateral trade

Bruno Arpino

Luca De Benedictis

Alessandra Mattei

RECSM Working Paper Number 44
October 2015

http://www.upf.edu/survey/_pdf/RECSM_wp044.pdf

Implementing Propensity Score Matching with Network Data: The effect of GATT on bilateral trade

Bruno Arpino
Department of Political and Social Science, Universitat Pompeu Fabra
E-mail: bruno.arpino@upf.edu

Luca De Benedictis
Dipartimento di Economia e Diritto, University of Macerata
E-mail: luca.debenedictis@unimc.it

Alessandra Mattei
Dipartimento di Statistica, Informatica, Applicazioni, University of Florence
E-mail: mattei@disia.unifi.it

Abstract

Motivated by the evaluation of the effect of GATT, we investigate the role of network information in propensity score matching. Under the assumption of strong ignorability, propensity score matching (PSM) is a widely used technique in causal inference studies to adjust for bias arising from an unbalanced distribution of observed confounders between a treatment and a control group. Both theoretical and applied works has recently considered the PSM for structured data, but the analysis of interlinked data is still missing. In this paper we consider the implementation of PSM in the context of network data. In our application, together with individual unit characteristics, also features of the social network in which units are embedded are considered as confounders (i.e., variables that impact on both the probability of receiving the treatment and the outcome). We study the sensibility of causal inference with respect to the presence of characteristics of the network in the set of confounders conditional on which strong ignorability is assumed to hold. We find that estimates of the average causal effect are sensitive to the presence of network information in the set of confounders, therefore we argue that estimates may suffer from omitted variable bias when network data are ignored, at least in our application.

Keywords: Centrality; Clustering; GATT; Matching; Networks; Trade; Unconfoundedness.

1 Introduction

The General Agreement on Tariffs and Trade (GATT) is a multilateral agreement aiming at regulating international trade among member countries by reducing tariffs and other trade barriers. It was signed in 1947 by 23 nations. Over the subsequent years, various rounds of negotiation took place and the number of countries joining the GATT has increased progressively to the actual number of 161. In 1995, after the conclusion of the Uruguay Round, the GATT was replaced by the World Trade Organization (WTO), a proper international organization with a new formal structure, a dispute settlement body, an extended agenda and new obligations for member countries (Irwin et al., 2008).

The impact of GATT/WTO membership on bilateral trade has been a subject of great interest in international business, economics and political science (e.g., Rose, 2004, 2007; Gowa and Kim, 2005; Subramanian and Wei, 2007; Tomz et al., 2007; Goldstein et al., 2007; Eicher and Henn, 2011). The approach usually adopted to evaluate the effect of GATT/WTO membership on bilateral trade is based on linear fixed effects regression methods, including either unit specific effects or unit and time specific effects (e.g., Rose, 2004; Tomz et al., 2007). Recently, Imai and Kim (2015), pointing out the limits of a causal interpretation of the estimates of linear fixed effects regression methods, have re-analysed data on GATT membership using innovative weighted fixed effects regression estimators, which the authors show to perform better in terms of bias than the standard fixed effects regression estimators.

In this paper we adopt an alternative approach framing causal inference in observational studies in the context of the potential outcome approach (Rubin, 1974, 1978), usually refereed to as Rubin's Causal Model (RCM; Holland, 1986).

In the potential outcome approach, observational studies have to be carefully designed to approximate randomized experiments for obtaining causal inferences that are objective (e.g., Rubin, 2008). Pre-treatment information is crucial to properly design observational studies. In some cases this information is difficult to obtain or the variables containing this information are unobservable. A common case is the one of social structure or social connections, in general, and of social network, in particular. If the social structure is of relevance, the position of a unit (an individual, a firm, a country) in the network, its direct and indirect connections, its tendency to cluster with its neighbours, all tend to convey some relevant information on latent characteristics of the unit. Disregarding this information could seriously bias the estimate of the causal effect of the treatment.

Recently, the availability of new data (adding on-line sources to the more traditionally used survey data), the development and consolidation of statistical techniques (Kolaczyk, 2009; Lusher et al., 2012) and the enlargement of the community of social network analysts to quantitatively based disciplines have fostered the statistical and econometric analysis of network data. Two streams of empirical research have gained momentum: the first one on (strategic) network formation (Robins et al., 2007; Butts, 2009; Christakis et al., 2010) and the second one, along Manski (1993) tradition, on peer effects when peers are connected through a group or a network structure (Lee, 2007; Graham, 2008; Bramoullé et al., 2009; Jackson et al., 2015). In spite of this new wave of research, very little has been done to properly frame the role of network structure in causal inference, and the issue of causality is still a very much debated one in network analysis (Doreian, 2001).

Here we discuss how to account for network data in causal inference by investigating the importance of using information on the network structure in the design phase of the GATT membership observational study.

The design of an observational study consists of two essential parts. The first part concerns the definition of causal effects as comparisons of potential outcomes under different treatments on a common set of units. In our study we consider as unit of analysis a country-dyad, that is, a pair of countries sharing a trade partnership. In order to clearly make the point, we do not consider the dynamic development of GATT, but we focus on the effect of GATT membership in 1954 on bilateral trade in 1955. Specifically, we focus on evaluating the average causal effect of GATT membership in 1954 of both countries against the aggregate of all possible alternatives. The choice of the period 1954-1955 is due to the relative stability in the number of countries included in the dataset, in the number of GATT members, and in the number of existing trade partnerships.

The second part of the design phase of the study concerns the explicit definition of an assignment mechanism, the stochastic rule that determines which units receive which treatment, that is, in our study, which dyads consist of two GATT members and which dyads consist of one or no GATT members. Specifically we need to introduce some assumptions on the assignment mechanism that allow us to draw inference on the causal effect of interest. Here we invoke the assumption of strong ignorability, which requires that the assignment mechanism is unconfounded, that is, free of dependence on potential outcomes, and probabilistic, that is, each dyad must have a positive probability of receiving either treatment (Rosenbaum and Rubin, 1983). In observational studies the plausibility of the strong ignorability assumption relies on the availability of information on all variables that may confound the relationship between the treatment and the outcome. In this paper, we consider the importance to include pre-treatment information on the network structure of the data in the set of variables conditional on which strong ignorability may be considered a reasonable assumption in our application. Formally, we assume that strong ignorability holds conditional on a set of pre-treatment variables that include both dyads' background characteristics and network information. Under this strong ignorability assumption we use propensity score matching to create a sub-sample of dyads where the distribution of the pre-treatment variables overlap and is well balanced between treatment groups. Given this sub-sample of dyads we move to the analysis phase, estimating the average causal effect of GATT membership on bilateral trade.

In order to clearly show the importance of using network information in the design phase of the study, we also investigate how the estimates of the average causal effect of GATT membership change when network information is ignored. We find that our estimates are sensitive to the presence of network information in the set of pre-treatment variables used to adjust treatment comparisons, suggesting that ignoring network information may lead to biased results.

The paper is organized as follows. In Section 2, we introduce the literature on trade policy evaluation, in general, and on the effect of the GATT/WTO membership on bilateral trade flows; we discuss and describe the data; and we define the world trade network and the statistics used to synthetically represent the level of connectivity in the structure of world trade. In Section 3, we describe the methodological framework we use. We first formally define the primitives for causal inference (units, treatments and potential outcomes) and the causal estimand of interest in our study. Then, we introduce the assignment mechanism, clearly specifying the assumptions we require for causal inference. Finally, we describe the role of propensity score in the design and analysis of our study. In Section 4, we discuss the results we find and Section 5 concludes.

2 International Trade and Trade Policy: Issue and data

The theory of international trade and commercial policy is one of the oldest branches of economic thought. Greek philosophers discussed the costs and benefits of opening up the home economy to competition of foreign traders and the very influential opinion of Adam Smith about the advantages of reducing tariffs on imported goods has been a milestone of the liberal thinking for centuries until today (Irwin, 1996; O'Rourke and Williamson, 2001). Until quite recently, the economic literature on the effects of trade policy was largely theoretical and, following Adam Smith, was massively concentrated on the effects of unilateral tariff reduction. With the emergence of regional processes of trade liberalization (e.g., the European Common Market, established in 1957, the ancestor of the European Union) and with the establishment of the GATT in 1947, the literature encompassed the study of the effects of preferential trade agreements at the regional and multilateral level. The classical model of international trade, under perfect competition and international price-taking countries, predicts that an increase in the country's trade volume (weighted by tariffs) is "a sufficient statistic for the total welfare effect" of trade liberalization (Baldwin and Venables, 1995). Empirically, this prediction has been interpreted at large as a simple rule: the gain from a Preferential Trade Agreement (PTA) is positively correlated with the increase in the import volume of country-members. This is, therefore, the theoretically based outcome variable to consider in a causal evaluation of the effect of a PTA, as the GATT.

2.1 The effect of GATT membership on trade

In spite of a clearly defined testable hypothesis, the empirical literature on the trade-effect of PTAs was dramatically sparse until the 2000s, and received a considerable boost when the seminal paper by Andrew K. Rose was published in 2004 (Rose, 2004). As previously mentioned, Rose (2004) examines the effect of GATT/WTO membership on bilateral trade flows by means of a linear fixed effects longitudinal regression model. The main result of the analysis is that the GATT has virtually no impact on trade flows: countries that are members of the GATT/WTO do not trade more than non-members countries.

This is a striking result. The GATT was established in 1947 having the explicit goal "...to remove or diminish barriers which impede the flow of international trade and to encourage by all available means the expansion of commerce" (Irwin, 1996). In particular, the GATT defined the rules governing trade policy and pursued a binding non-discriminatory tariff-reduction multilateral strategy, in which the (best) concessions between any two country-members were automatically passed to other members according to the most-favoured-nation principle. All formal members (as classified by Rose, 2004) had to fulfil GATT's rules, but those rules applied also to non-member participants (as classified in Tomz et al., 2007): Colonies and overseas territories (Art. XXVI of GATT), newly independent states, and provisional members.

In his estimates of the effect on international trade of GATT/WTO membership, Rose (2004) classifies only the formal members of GATT as "members", and in a standard (reduced form) gravity model of international trade (De Benedictis and Taglioni, 2011; Anderson, 2011; Head and Mayer, 2014) he uses a country specific fixed effect estimator for panel data, controlling for a comprehensive set of time fixed effects, and a large set of covariates. Rose (2004) builds a dataset covering bilateral trade flows for 178×177 country pairs, between 1948 and 1999. Countries that do not report any trade flows are excluded from the dataset. The outcome variable is a measure of yearly bilateral trade: the average value of bilateral trade between a pair of countries, i and j, considering all measures potentially available. To avoid misreported trade flows, all the four possible flows are used to calculate the average: exports from i to j, imports into j from i, exports from j to i, imports into i from j, previously deflated by the US Consumer Price Index. The covariates are variables now considered standard in the gravity model literature: the real gross domestic product (GDP) of country i and country j, included as the logarithm of the product; the logarithm of the bilateral distance between i and j, measured using the great-circle distance formula, and a dummy variable indicating if country i and country j are sharing a land border. Other variables augmented the gravity model in order to reduce the possibility of an omitted variable bias: the log product of real GDP per capita; the log product of land area; the number of island nations in the country-pair (0, 1, or 2); the number of landlocked nations in the country-pair (0, 1, or 2); a dummy variable equal to one if i and j share a common language and zero otherwise; a dummy variable equal to one if i and j were ever colonies after 1945 with the same colonizer; a dummy variable equal to one if i is a colony of j at time t or vice versa; a dummy variable equal to one if i ever colonized j or vice versa; a dummy variable equal to one if i and j belong to a Currency Union or share the same currency at time t; a dummy variable equal to if i and j are part of the same nation during the sample period; a dummy variable equal to one if i and j both belong to the same regional trade agreement; a dummy variable equal to one if country i is offering a Generalized System of Preferences (GSP) to country j or vice versa at time t (see Rose (2004), Tomz et al. (2007) for a detailed description of the dataset and the source of each variable). The last three variables are omitted in our analysis because of lack of variation in 1954.

The main independent variables of interest in Rose (2004) are a binary variable which is equal to one if both i and j are GATT/WTO members at t, and a binary variable which is equal to one if either i or j is a GATT/WTO member at t. The bottom line of Rose (2004) is that " ...the GATT/WTO does not seem to have had much of an impact on trade".

This thought-provoking result encouraged the research on the topic and produced a series of contributions along the same methodological line (Rose, 2007; Gowa and Kim, 2005; Subramanian and Wei, 2007; Tomz et al., 2007; Goldstein et al., 2007; Liu, 2009; Eicher and Henn, 2011) that, as summarized by Maggi (2014), "...overturned or qualified Rose's results."

Our analysis stands on the shoulder of these contributions. To be consistent with the original set of information, we take the Rose (2004) data as it is, disregarding the well posed critiques by Subramanian and Wei (2007) on the use of an average of trade flows as the relevant outcome variable (artificially generating a symmetry between i and j), or by Liu (2009) on the implicit sample selection generated by the logarithmic transformation of the outcome variable, implying the exclusion of zero-trade flows from the analysis. Our analysis is especially related to Tomz et al. (2007), adopting their revised classification of GATT members (also used in Imai and Kim, 2015) and considering the set of covariates used both in Rose (2004) and Tomz et al. (2007). Eicher and Henn (2011) is a second contribution especially related to our own analysis. In their evaluation of the trade effect of the GATT/WTO, they include in the gravity equation a fundamental element suggested by the modern theory of multilateral trade liberalization based of the role of terms-of-trade externalities (Bagwell and Staiger, 2004), namely, they account for the fact that countries with higher market power would be advantaged by a multilateral tariff reduction more than countries with lower market power. Therefore, a measure of market power is included in the set of covariates to avoid a potentially strong omitted variable bias. In our analysis, network variables can be interpreted as proxies of market power.

Our analysis also owes some credit to Baier and Bergstrand (2009), which are the first to apply propensity score matching techniques to the analysis of the effect of trade policy on trade flows.

2.2 Trade as a network

In recent years a series of contributions have analysed the structure of world trade using the tools of social network analysis (see De Benedictis and Tajoli, 2011; De Benedictis et al., 2014; Fagiolo, 2015, for a recent review of the literature). All of them share the implicit assumption that international trade between two agents of country i and country j is the result of a complex choice depending on the characteristics of i

and j and on several elements whose nature is dyadic (e.g., bilateral distance), as in the gravity model of international trade. The opportunity cost of trading generates an externality that is only observable in its final realization, which is the structure of the world trade network.

Given the characteristics of the data built by Rose (2004) and Tomz et al. (2007), the world trade network is an undirected graph $G(\mathcal{N}, \mathcal{E})$ that consists of a set of nodes, identifying potential trade partners $\mathcal{N} = \{1, \ldots, n\}$, and a list of unordered pairs of nodes, called edges, that correspond to realized trade partnerships: $\mathcal{E} = \{(i,j), (k,l), (j,l), \ldots\}$ for i, j, k, and $l \in \mathcal{N}$. The visualization of G in 1948 (G^{1948}), used only for explicative purposes, is included in panel (a) of Figure 1. In that case, the dimension of the graph is $G^{1948}(11, 54)$, with the 11 reported countries being identified by their ISO3 character codes, so that $\mathcal{E} = \{(USA, GBR), (CAN, CHL), \ldots\}$. While the number of realized partnerships is 54, the number of potential partnerships is 55 (a.k.a., $\frac{11 \times 10}{2}$), making the density of G^{1948} equal to $\delta^{1948} = \frac{54}{55} = 0.982$, with no country being an isolate.

G can be also expressed in matrix form by its adjacency matrix $\mathcal{L} = [\mathcal{L}_{ij}]$ where

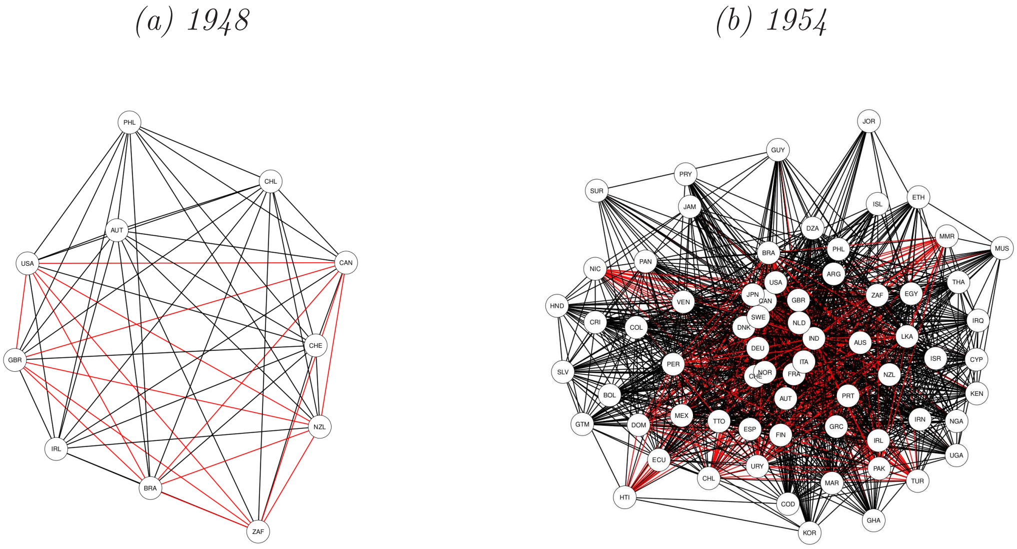
$$\mathcal{L}_{ij} = \begin{cases} 1 & \text{if } \{i,j\} \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

The red edges in Figure 1 indicate that both countries at the extremes of an edge are members of the GATT in the corresponding year, according to Tomz et al. (2007), black edges indicate that at least one of the two countries is not a GATT member.

In 1954, the dimension of the graph is $G^{1954}(66, 1319)$ and the corresponding density is $\delta^{1954} = \frac{1319}{2145} = 0.615$. What is evident from Figure 1 is that the number of GATT members increases substantially and that the countries' position in the network changes over time. The world trade network, as represented by the data, was symmetric in 1948, with no country playing a central role in the structure of world trade, while it assumes a core-periphery structure in 1954, with some countries assuming a central position while others lagging at the boundaries of the network.

Network analysis provides several indicators to assess the importance of a node centrality, capturing different aspects of its position. Here we concentrate on two

Figure 1: Trade Partners in 1948 and 1954



(a) 1948

(b) 1954

Note: The figure represents the graph $G(\mathcal{N}, \mathcal{E})$ of trade partners in 1948 (panel (a)), and 1954 (panel (b)). Countries (nodes) are identified by their ISO3 character code. Red edges indicate that both countries are members of the GATT in the corresponding year, according to Tomz et al. (2007). Dimensions of the graphs are: $G^{1948}(11, 54)$ and $G^{1954}(66, 1319)$. Countries that do not report any trade flows are excluded from the dataset. Elaborations are our own. Further description of the data and of the data sources is in Rose (2004) and Tomz et al. (2007).

standard measures of centrality (Wasserman and Faust, 1994): the degree centrality, C_D, and the eigenvector centrality measure, C_E. The degree centrality measures how a node is connected to others and it is a measure of local centrality because it considers only the direct neighbours. The eigenvector centrality measure is an index based on direct and indirect neighbours' characteristics, measuring how important, central, influential or tightly clustered a node's neighbours are. The eigenvector centrality is therefore a measure of global centrality (see De Benedictis et al., 2014, on the application of these measures to international trade flows).

Degree centrality is the simplest measure of the position of a node in a network. Since the trade network is considered in its binary version, C_D measures the centrality of a node by the number of connections the node has. Formally,

$$C_{Di} = \sharp\{j : \mathcal{L}_{ij} = 1\} = \sum_{j=1, j\neq i}^{n} \mathcal{L}_{ij}, \tag{1}$$

where $\sharp$ denotes the cardinality of a set.

Let's recall that n is the total number of nodes (countries) in the network, and $\mathcal{L}_{ij}$ is the element (i, j) in the equivalent trade adjacency matrix $\mathcal{L}$, where i is the row-indicator corresponding to exporting countries, and j is the column-indicator corresponding to importing countries. If $\mathcal{L}_{ij} = 1$ the two countries i and j are trade partners (regardless of the direction of trade flow), if $\mathcal{L}_{ij} = 0$ they are not trading between them.

The degree centrality measure, being dependent on the number of the existing nodes in the network, makes it difficult to compare networks of different number of node. It is usually better to calculate the normalized version of C_D using the total number of possible neighbours excluding self, $n - 1$, or the maximum degree centrality realized among the nodes, $\max \sharp\{j : \mathcal{L}_{ij} = 1\}$, as a normalizing factor. We opted for this second possibility:

$$\tilde{C}_{Di} = \frac{\sum_{j=1,j\neq i}^{n} \mathcal{L}_{ij}}{\max \sum_{j=1,j\neq i}^{n} \mathcal{L}_{ij}}. \tag{2}$$

It follows that this indicator ranges from 0 to 1; the more its degree centrality is close to 1, the more a country is connected in the network. As an example, Chile (CHL) has $\tilde{C}_{D\,CHL} = 1$ in 1948, and $\tilde{C}_{D\,CHL} = 0.661$ in 1954; while Great Britain (GBR) has $\tilde{C}_{D\,GBR} = 1$ in both years.

A measure of global centrality is the eigenvector centrality, in which the position of each node's neighbours, and the direct and indirect links of a node, all contribute in determining the centrality of a node. The eigenvector centrality is determined by the eigenvector centrality of its neighbours: It is not the country's centrality itself that matters, what really matters is the centrality of the countries linked to it. The circularity of the argument is evident, but can be tackled using some matrix algebra.

Starting from a binary trade-matrix, it is possible to define the eigenvector centrality of country i as the sum of the eigenvector centralities of its neighbours. That is:

$$C_{Ei} = \mathcal{L}_{i1}C_{E1} + \mathcal{L}_{i2}C_{E2} + \cdots + \mathcal{L}_{i(n-1)}C_{En-1} + \mathcal{L}_{iN}C_{En}. \tag{3}$$

The system of equations for the eigenvector centrality of all n countries can be rewritten in matrix form as:

$$(I - \mathcal{L})\overrightarrow{C_E} = 0, \tag{4}$$

where I is a $n \times n$ identity matrix, $\mathcal{L}$ is the trade adjacency matrix, and $\overrightarrow{C_E}$ is the $n \times 1$ vector of countries' eigenvector centralities (Equation 4 is the trade-matrix characteristic equation for an eigenvalue λ=1). From the Perron-Frobenius theorem we know that a square matrix with positive (and some classes of non-negative) real entries has a unique largest real eigenvalue and that the corresponding eigenvector has strictly positive components. Adopting a proper row-normalization (but in our case, given the symmetry of the trade data, also a column-normalization will give the same result), we can consider the entries of the relevant (main) eigenvector as a measure of country centrality.

As in the case of the degree centrality, we normalize the eigenvector centrality using the maximum centrality:

$$\tilde{C}_{Ei} = \frac{C_{Ei}}{\max C_{Ei}}. \tag{5}$$

Also this indicator ranges from 0 to 1. As an example, Chile (CHL) has $\tilde{C}_{E\,CHL} = 1$ in 1948, and $\tilde{C}_{E\,CHL} = 0.746$ in 1954; while Great Britain (GBR), being connected to all central trade partners, has $\tilde{C}_{E\,GBR} = 1$ in both years.

In general, a high level of $\tilde{C}_E$ corresponds to countries belonging to large and cohesive (high-density) sub-networks, or, in other terms, countries with a high value of $\tilde{C}_E$ are connected to many other countries which are, in turn, connected to many others.

This property can be better qualified by high order network statistics (Wasserman and Faust, 1994). We consider two of them, the first one belonging to a class of local clustering measures and the second one to a class of global clustering measures.

The local clustering, or local transitivity, is the probability that the adjacent nodes of a node i (the neighbourhood of country i, $\mathcal{N}_i$) are connected. In the case of an undirected graph, the local transitivity is the ratio of the triangles connected to the node and the triples centred on the node (Wasserman and Faust, 1994).

More formally, following Watts and Strogatz (1998), if a node i has C_D neighbours, the number of possible edges that could exist among the nodes within the neighbourhood is $\frac{C_{Di}(C_{Di}-1)}{2}$. Thus, the local clustering coefficient for country i can be defined as:

$$C_{Ti} = \frac{2\sharp\{(j,k) : j,k \in \mathcal{N}_i, (j,k) \in \mathcal{E}\}}{C_{Di}(C_{Di} - 1)}. \tag{6}$$

This measure is equal to 1 if every neighbour connected to i is also connected to every other node within the neighbourhood, and 0 if no node that is connected to i connects to any other node that is connected to i.

The measure of global clustering that we consider, denoted by C_C, is generated using a "community detection algorithm" (Newman, 2006) and in particular the Spin Glass algorithm of Reichardt and Bornholdt (2006). For every graph $G(\mathcal{N}, \mathcal{E})$, this algorithm separates dense sub-graphs, also called communities, via a spin-glass model and simulated annealing. A community is, therefore, a set of nodes with many edges inside the community and fewer edges outside it. In our application, the algorithm classifies countries into three classes.

All these measures, $\tilde{C}_D; \tilde{C}_E; C_T; C_C$, define for each country i in the world trade network its relative position and its degree of connectivity at the local and the global level. In the next sections we will investigate how much those measures are relevant for estimating causal effects of GATT membership on bilateral trade flows.

3 Methodological Framework

3.1 Basic setup

In estimating the causal effect of GATT, we adopt the potential outcome approach to causal inference (e.g., Rubin, 1974, 1977, 1990; Imbens and Rubin, 2015), which is often referred to as Rubin's Causal Model (RCM, Holland, 1986). To draw inference on causal effects under the RCM, we first need to introduce and define the basic concepts for causal inference, that is, units, treatments and potential outcomes. In our study units are country-dyads in $t = 1954$, the year we focus on.

The treatment variable of interest is belonging to the GATT in 1954, as a formal member or as non-member participants (as classified in Tomz et al., 2007). Analytically, let T_{ij} be the treatment variable. T_{ij} is a binary variable equal to one if both countries in dyad (i, j) are members of GATT in 1954 and zero otherwise. We are interested in assessing the effect of GATT membership in 1954 on bilateral trade in 1955, which is measured by the logarithm of the average trade flows for each dyad, as previously defined. We allow for a one-year lag when considering the outcome variable to guarantee that the treatment is measured before the outcome and not simultaneously. For each dyad (i, j), we can define two potential outcomes for bilateral trade, $Y_{ij}(0)$ and $Y_{ij}(1)$, which are, respectively, the value of bilateral trade in 1955 for dyad (i, j) if at least one of the countries in the dyad is not members of GATT in 1954, and the value of bilateral trade in 1955 for dyad (i, j) if both countries in the dyad are members of GATT in 1954.

The fact that only two potential outcomes for each dyad are defined reflects the acceptance of the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1990), which rules out hidden versions of treatments and interference between units. The no hidden versions of treatments component of SUTVA implies that each level of the treatment defines a single outcome for each dyad. In our setting, this assumption might be arguable because a dyad that is assigned to treatment level zero can receive different forms of the treatment, that is, the dyad may comprise either only one GATT member or no GATT member, and so $Y_{ij}(0)$ can be unstable. Nevertheless, it is reasonable to believe that GATT membership can affect bilateral trade only when countries mutually agree on reducing trade barriers, and so the assumption of no hidden versions of treatments can be viewed as a reasonable assumption. Anyway, our approach could be easily extended to the case of multivalued treatments.

The no interference assumption implies that a dyad's trade volume is not affected by other dyads' GATT membership status. Although this assumption may be untenable in a context of world trade network, we maintain it in the present analysis. Our primary goal is to investigate the role of pre-treatment information on network characteristics in the design and analysis of an observational study. Extensions permitting interference among units is, however, at the top of our research agenda.

The causal estimand we focus on is the population average treatment effect (ATE), which is defined as the mean difference between potential outcomes:

$$ATE = \mathbb{E}[Y_{ij}(1) - Y_{ij}(0)]. \tag{7}$$

Unfortunately, we can never observe both potential outcomes for each dyad, instead we can only observe one of the potential outcomes for each dyad, either $Y_{ij}(0)$ or $Y_{ij}(1)$, depending on the treatment actually received (a.k.a., the fundamental problem of causal inference; Holland, 1986; Rubin, 1978). Let Y_{ij}^{obs} be the observed trade volume for dyad (i, j), then $Y_{ij}^{\text{obs}} = T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0)$. For each dyad (i, j), we also observe the treatment actually received, T_{ij}, and a set of pre-treatment background characteristics, $\boldsymbol{X}_{ij}$. The vector of the observed covariates, $\boldsymbol{X}_{ij}$, includes both dyad-specific information, such as the indicator for common language and the indicator for currency union, as well as dyad information based on country-specific characteristics, such as the sum of the logarithms of the real GDP for country i and country j (see Section 2.1 for a list of the pre-treament variables we use in the analysis). Unfortunately, we do not have country-specific information in the original data of Rose (2004) and Tomz et al. (2007), but it is worth noting that country-specific characteristics might be valuable in the design phase of the study. Finally, the dyadic structure of the data naturally provides information on pre-treatment trade relationships. In our study we summarize the network structure of the data using two measures of network centrality (degree centrality and eigenvector centrality), the local clustering coefficient and the set of indicators for clusters of nodes derived from community detection algorithms. Let $\boldsymbol{N}_i = \left[\tilde{C}_{Di}, \tilde{C}_{Ei}, C_{Ti}, C_{Ci}\right]$ denote the vector of network measures for country i, $i = 1, \ldots, n$.

3.2 The Assignment Mechanism

Carefully designing a study is crucial for drawing objective inferences for causal effects (Rubin, 2008). An essential part of the design-phase of a study concerns the explicit specification of the treatment assignment mechanism, the stochastic rule that determines which units receive which treatments, and so which potential outcomes are realized and which are missing. In observational studies the treatment assignment mechanism is usually unknown and we need to posit one, introducing plausible assumptions.

We invoke the assumption that the treatment assignment mechanism is strongly ignorable given pre-treatment variables and network measures, adapting to the case of an observational study with network data the original definition given by Rosenbaum and Rubin (1983). Formally

Assumption 1 *(Strong ignorability given pre-treatment variables and network measures) The treatment assignment mechanism is strongly ignorable if the following conditions hold:*

Unconfoundness: $T_{ij} \perp (Y_{ij}(0), Y_{ij}(1)) \, | \boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j$

Overlap: $0 < P(T_{ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j) < 1.$

Unconfoundness amounts to assuming that the treatment assignment is independent of the potential outcomes conditional on the observed covariates and network measures. The overlap assumption imposes that the treatment assignment is probabilistic implying that there is sufficient overlap in the joint distribution of the covariates and network measures between treated and control dyads. In other words, strong ignorability implies that within cells defined by the values of pre-treatment variables and network measures, the treatment is as randomly assigned, so that the comparison of treated and control dyads with the same value of the observed covariates and network measures leads to valid inference on causal effects. In the literature, strong ignorability is usually assumed conditional on the observed background characteristics only, and network information is not used, either because it is not available or because it is ignored. In our study the network structure is reasonably correlated both with potential outcomes for bilateral trade as well as with GATT participation, therefore ignoring it may induce bias.

3.3 Propensity Score Matching with Network Data

Under Assumption 1, we can remove all biases in comparisons between treated and control dyads by adjusting for differences in observed covariates and network measures. Although feasible in principle, in practice this will be difficult to implement with a large number of background variables. We can deal with the curse of dimensionality using the propensity score, defined as the unit-level probability of receiving the treatment given the observed covariates (Rosenbaum and Rubin, 1983). In our study the propensity score is defined as the conditional probability that both countries in a dyad are GATT members given the observed background covariates and network measures:

$$e_{ij} = P(T_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j). \tag{8}$$

Rosenbaum and Rubin (1983) show that the propensity score has two key properties, which can be re-formulated as follows in our setting. The propensity score is a balancing score, that is, the treatment is independent of pre-treatment background covariates and network measures given the propensity score:

$$T_{ij} \perp (\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j)\,|e_{ij};$$

and if assignment to treatment is strongly ignorable given pre-treatment background variables and network measures, then assignment to treatment is strongly ignorable given the propensity score, that is, if

$$T_{ij} \perp (Y_{ij}(0), Y_{ij}(1))\,|\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j \qquad \text{and} \qquad 0 < P(T_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j) < 1$$

then,

$$T_{ij} \perp (Y_{ij}(0), Y_{ij}(1))\,|e_{ij} \qquad \text{and} \qquad 0 < P(T_{ij} = 1|e_{ij}) < 1.$$

These properties imply that adjusting for differences in the propensity score is sufficient for removing the bias associated with differences in the pre-treatment background variables and network measures.

As in barely all observational studies, we do not know the true propensity score, so we need to estimate it. We specify a logit regression model for the propensity score:

$$e_{ij} = \frac{\exp\{g\,(\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j; \alpha)\}}{1 + \exp\{g\,(\boldsymbol{X}_{ij}, \boldsymbol{N}_i, \boldsymbol{N}_j; \alpha)\}},$$

where α is a parameter vector and g is a function of all the covariates, including both dyads' characteristics and network measures.

The estimation of the propensity score requires some effort to find a function g of all the covariates that leads to estimates of the propensity score that balance the covariate distributions between treatment and control dyads in the sample. Typically, this implies including higher order terms for some of the covariates and/or interactions among covariates (Dehejia and Wahba, 1999).

Once the propensity score is estimated, several methods of matching are available. The most common ones are kernel, nearest neighbour, radius, and caliper (for a discussion about these methods see, e.g., Smith and Todd, 2005; Caliendo and Kopeinig, 2008; Stuart, 2010).

Similar to the specification of the propensity score, also the choice of the matching algorithm has to be guided by the goal of maximizing the balance of the covariates. Following common practice, we assess the balance of each covariate using the Absolute Standardized Bias (ASB). The ASB for a given variable X is defined as the absolute value of the difference of means between treatment and control group standardized by the average standard deviation in the treatment and control groups (Rosenbaum and Rubin, 1985). The ASB is a measure of covariate balance: a lower ASB indicates that the treatment and control groups are more similar with respect to the given covariate. The process of adjusting the specification of the propensity score and choosing among the matching algorithm should be stopped only when an acceptable balance solution is achieved. Although the goal of matching is to eliminate any imbalance in observed covariates, pragmatically in the common practice an ASB smaller than 10% for each covariate is usually considered as acceptable (Normand et al., 2001). We follow this practice and choose kernel matching because it guaranteed the best balance among the numerous solutions we tried.

The basic idea of matching for estimating ATE is that for each dyad a set of dyads should be found in the opposite treatment group that are sufficiently close to it. In propensity score matching the distance metric is defined by the propensity score. The unobserved potential outcome for each dyad is then estimated as the average observed outcome for the matched dyads and ATE is estimated averaging over all dyads that found at least one matched dyad in the opposite treatment group. Formally, let I_1 and I_0 denote the set of treated and control dyads, respectively. Let tt' and cc' indicate a generic treated and control dyad, respectively, and let $A_{tt'}$ denote the set of the indices of control dyads matched to $tt' \in I_1$. Similarly, $A_{cc'}$ indicates the set of the indices of treated dyads matched to $cc' \in I_0$. After $A_{tt'}$ and $A_{cc'}$ have been constructed for all units, the matched dataset M is built selecting all treated (control) dyads that successfully found a matched control (treated) dyad and all matched control (treated) dyads, while all other dyads are discarded. Formally,

$$M = \{tt' : A_{tt'} \neq \emptyset\} \cup \{\cup_{tt'} A_{tt'}\} \cup \{cc' : A_{cc'} \neq \emptyset\} \cup \{\cup_{cc'} A_{cc'}\}\,. \tag{9}$$

Asymptotically, all PSM estimators should converge to the same results (Smith, 2000), while in small samples the choice of the matching algorithm can be important and generally a trade-off between bias and variance arises (Caliendo and Kopeinig, 2008). The different matching algorithms differ in the way the sets $A_{tt'}$ and $A_{cc'}$ and weights attributed to each matched unit within them are defined. In principle, kernel matching (Heckman et al., 1998) consists of matching each treated (control) dyad with all the available control (treated) dyads, i.e., $A_{tt'} = I_0$ and $A_{cc'} = I_1$, with weights depending on the distance in terms of propensity scores between each treated and control dyad. Usually, symmetric, nonnegative and unimodal kernel functions are used so that higher weights are placed on dyads close in terms of the propensity score of a matched dyad and lower weights on more distant dyads (Heckman et al., 1997). A general expression for weights is:

$$w(tt', cc') = w(cc', tt') = \frac{K\left(\dfrac{\hat{e}_{tt'} - \hat{e}_{cc'}}{h}\right)}{\sum K\left(\dfrac{\hat{e}_{tt'} - \hat{e}_{cc'}}{h}\right)}, \tag{10}$$

where h indicates the bandwidth.

The exact definition of the weights depends on the specific kernel function that it is used. It has been shown that the choice of the kernel function makes little difference in practice (Di Nardo and Tobias, 2001). We use the Epanechnikov kernel, so that weights are defined as follows:

$$K(u) = \begin{cases} \dfrac{3}{4}(1-u^2) & \text{if } |\hat{e}_{tt'} - \hat{e}_{cc'}| \leq h; \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where we set $h = 0.06$.

In words, this type of kernel retains dyads that do not differ more than h units of propensity score. Formally, this algorithm defines the set of matched controls for each treated dyad as follows:

$$A_{tt'} = \{cc' \in I_0 : |\hat{e}_{tt'} - \hat{e}_{cc'}| \leq h\}\,. \tag{12}$$

Similarly, the set of matched treated for each control dyad is:

$$A_{cc'} = \{tt' \in I_1 : |\hat{e}_{cc'} - \hat{e}_{tt'}| \leq h\} . \tag{13}$$

The ATE estimator takes the form:

$$\begin{aligned}\widehat{ATE} =& \frac{1}{\sharp(I_1 \cap M)} \left\{ \sum_{tt' \in I_1 \cap M} \left(Y_{tt'} - \sum_{cc' \in A_{tt'}} Y_{cc'} w(tt', cc') \right) \right\} \\ &+ \frac{1}{\sharp(I_0 \cap M)} \left\{ \sum_{cc' \in I_0 \cap M} \left(Y_{cc'} - \sum_{tt' \in A_{cc'}} Y_{tt'} w(cc', tt') \right) \right\} ,\end{aligned} \tag{14}$$

where $I_1 \cap M$ represents the subset of treated dyads that found at least one matched control dyad and $I_0 \cap M$ represents the subset of control dyads that found at least one matched treated dyad.

Note that, before implementing the matching, dyads with extreme estimated propensity scores are discarded to satisfy the balancing property. Following the consolidated practice we retain dyads with propensity scores falling in the interval where the propensity score distributions of treated and control dyads overlap.

4 Results

As said in the previous section, we tried different specifications of the propensity score combined with different matching algorithms in order to obtain a good balance in each of the covariate used to estimate the propensity score. We present results from the best solution we obtain for three different cases that differ with respect to the set of variables included in the propensity score estimation:

1. $\boldsymbol{X}$ only;

2. $\boldsymbol{X}$ plus $\tilde{C}_D$ (degree centrality);

3. $\boldsymbol{X}$ plus $\boldsymbol{N}$ (all the network variables).

In the first case, we only include the background covariates, $\boldsymbol{X}$, in the propensity score model, omitting network variables. We consider this case to assess the possible bias that we would obtain under Assumption 1 when network variables are ignored.

In this case, we choose the propensity score matching solution that gives the best balance of the $\boldsymbol{X}$ variables, the only variables used in estimating the propensity score. In fact, this would be the procedure that a researcher that ignores network variables would follow. However, we report also balance of the $\boldsymbol{N}$ variables to show to what extent the misspecified propensity score reduces imbalance also on the network variables, $\boldsymbol{N}$, that are not used in its estimation.

As a second case, we consider including in the set of independent variables of the propensity score model only one network variable, namely, degree centrality of each of the two countries in a dyad, together with the $\boldsymbol{X}$ variables. Also in this case we show the balance obtained on all the covariates. In this case, we choose the solution that gives the best balance of the $\boldsymbol{X}$ variables and degree centrality measures. Under Assumption 1, ATE estimators based on a propensity score that only includes degree centrality as a network measure would still be biased. However, we aim at assessing whether all network variables are better balanced compared to the first case even if only one network measure is included in the propensity score model. In other words our goal is to determine to what extent adjusting the propensity score for the simplest network measure would also improve the balance of the other network measures and consequently assess the robustness of ATE estimate to the inclusion of only one network measure instead of all of them.

Finally, we consider estimation of the propensity score with all the $\boldsymbol{X}$ and $\boldsymbol{N}$ variables. In this case, we select the propensity score matching solution that guarantees a good balance of all variables ($\boldsymbol{X}$ and $\boldsymbol{N}$).

Note that, in addition to the main effects of each variable, the three specifications of the propensity score also include some higher order terms and interactions.

We start presenting the results on the balance of covariates obtained in each of the three cases. For each of the variable we calculate the percent ASB before and after matching. The percent ASB for each variable is presented in Table A.1 in the appendix. In Table 1 we summarize this information by presenting the mean and median percent ASB calculated on the $\boldsymbol{X}$ variables only, on the $\boldsymbol{N}$ variables only, and on all variables ($\boldsymbol{X}$ and $\boldsymbol{N}$) under the three specifications of the propensity score.

Table 1: Mean and median Percent Absolute Standardized Bias for $\boldsymbol{X}$ and $\boldsymbol{N}$ separately; and for the joint set of $\boldsymbol{X}$ and $\boldsymbol{N}$, under three different specifications of the propensity score

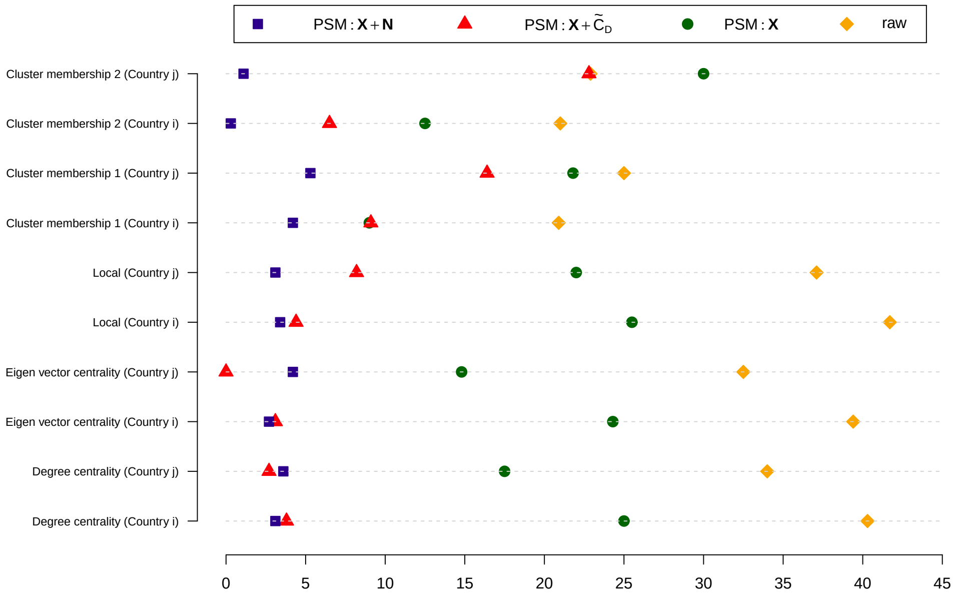
		The propensity score model includes:					
		$\boldsymbol{X}$		$\boldsymbol{X}$ *and* $\tilde{C}_D$		$\boldsymbol{X}$ *and* $\boldsymbol{N}$	
		Mean	*Median*	*Mean*	*Median*	*Mean*	*Median*
ASB% of	*Sample*	*ASB%*	*ASB%*	*ASB%*	*ASB%*	*ASB%*	*ASB%*
$\boldsymbol{X}$	Raw	18.8	17.2	18.8	17.2	18.8	17.2
	Matched	1.2	0.6	2.4	2.0	2.5	2.4
$\boldsymbol{N}$	Raw	31.5	33.3	31.5	33.3	31.5	33.3
	Matched	20.2	21.9	7.7	5.5	3.1	3.3
$\boldsymbol{X}$ and $\boldsymbol{N}$	Raw	24.6	22.4	24.6	22.4	24.6	22.4
	Matched	10.1	3.9	4.8	3.3	2.8	3.1

Let first consider the balance of $\boldsymbol{X}$ variables. Table 1 shows that in the raw dataset there is a substantial imbalance in the distribution of covariates. In fact, the mean and median percent ASB are respectively 18.8% and 17.2%. From Table A.1 in the appendix we can see that the ASB for several covariates is well above the threshold of 10%. The highest ASB is for the logarithm of the product of real GDP (39.1%). This means that a raw comparison between the outcomes of treated and untreated dyads is likely to give a biased estimate of ATE because confounded with the effect of the $\boldsymbol{X}$ variables. The three propensity score procedures aim at adjusting treatment comparisons for these differences. We can judge to what extent possible biases due to confounders are reduced by considering ASB after matching. All the three propensity score matching procedures succeed in drastically improving the balance of $\boldsymbol{X}$ variables. The mean percent ASB for the $\boldsymbol{X}$ variables drops from 18.8% to 1.2% with the first propensity score matching approach and to about 2.5% in the other two cases. However, the balance of $\boldsymbol{X}$ variables remains very good in all cases (the maximum percent ASB is 6.6% for one covariate in the third case, while for several covariates is much smaller).

What about the balance of $\boldsymbol{N}$ variables? In Figure 2, we compare the performance of the three matching approaches in terms of reduction in the ASB of each of the network variables as compared to the raw (unmatched) dataset. Figure 2 shows that the distribution of the network measures across the groups of treated and control dyads are quite different before matching: the percent ASB ranges from 21% to 41.7% (see also Table A.1). Table 1 indicates that the mean and median percent ASB for $\boldsymbol{N}$ are, respectively, 31.5% and 33.3% before matching. As a consequence, when considering the balance of both $\boldsymbol{X}$ and $\boldsymbol{N}$ in Table 1 the mean and median percent ASB increase.

The first propensity score matching approach, that ignores network variables, as expected, does not behave well in terms of balance of $\boldsymbol{N}$. In fact, Figure 2 shows

Figure 2: Percent absolute standardized bias of network variables with three different specifications of the propensity score.



Note: In each row the graph reports the % ASB for one of the network measures obtained with three different specifications of the propensity score: including all variables (PSM: $\boldsymbol{X}$ and $\boldsymbol{N}$), including $\boldsymbol{X}$ variable and degree centrality ($\boldsymbol{X}$ and $\tilde{C}_D$) and including only $\boldsymbol{X}$ variables. It is also reported the ASB calculated on the raw (unmatched) dataset.

Table 2: Dyads on and off common support by treatment status with three different specifications of the propensity score

	The propensity score model includes:						
	$\boldsymbol{X}$		$\boldsymbol{X}$ *and* $\tilde{C}_D$		$\boldsymbol{X}$ *and* $\boldsymbol{N}$		
	Common support		*Common support*		*Common support*		
Treatment group	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Total*
Untreated	0	766	11	755	22	744	766
Treated	5	548	42	511	48	505	563
Total	5	1314	53	1266	70	1249	1319

that although there is a reduction in the ASB for each of the network variables with respect to the raw dataset, the ASB remains quite high (between 20% and 30% for most of these variables). The second propensity score matching approach that only adds degree centrality improves considerably the balance with respect to the naive approach that excludes all $\boldsymbol{N}$ variables. In this case, in fact, the percent ASB of most of the network variables is below 10% with an average of 7.7%. Therefore, in our dataset including only the simplest measure of network is sufficient to obtain a quite good balance also on other network variables. However, Figure 2 shows that for two global cluster membership indicators the percent ASB remains high. Figure 2 shows that the third propensity score matching approach where all $\boldsymbol{N}$ variables are used further reduces the imbalance of the network variables (maximum ASB = 5.3%; average ASB = 2.8%).

Table 2 reports the number of treated and control dyads that are on and off the common support of the propensity score distribution. As said in the previous section, dyads whose propensity score are outside the range of the propensity score distribution in the opposite treatment group are discarded. Moreover we have already noticed that dyads that do not find a match in the opposite group within the bandwidth are also excluded in the Epanechnikov kernel matching. Table 2 shows that by expanding the set of variables included in the propensity score matching, the number of unmatched dyads increases from 5 to 70, i.e., from 0.4% to 5% of the total sample size. In other words, the first propensity score matching ignores the fact that some dyads in the two treatment groups may be too different in terms of their network characteristics. The better balance of the network variables obtained in the second and third propensity score matching approaches may be in part due to discarding these “incomparable” dyads.

In any empirical causal inference study it is important to investigate the characteristics of the unmatched units. In fact, it has to be recognized that discarding unmatched units from the analyses may change the estimand (Crump et al., 2009). This is because the unmatched units may hold particular characteristics that make impossible to compare them with units in the opposite treatment group without relying on model extrapolations. In Table 3 we report the means of each covariate and network measure for treated and control dyads, separately, and further distinguishing within each group dyads that found a match in the opposite group (“on support”, denoted by ON in Table 3) and those that did not (“off support” denoted by OFF in Table 3). From this table we can get several interesting insights. First, for some covariates, dyads off support have quite different average values than dyads on support in the same treatment group. For example, using the first propensity score model that only includes $\boldsymbol{X}$, 60% of unmatched treated dyads are currently in a colonial relationship compared to 1.3% of matched treated dyads. Second, we notice that the group of covariates that show quite different average values for dyads on and off support remain stable for the three propensity score models. These covariates include the currency union indicator, variables related to colonial relationships and common language. Third, it is interesting to notice that propensity score models that include network variables help detecting important differences between treated and control dyads with respect to network characteristics. Both degree and eigenvector centrality measures are on average much lower among unmatched control dyads as compared to control dyads that succeed in finding a match in the treatment group. This indicates that it is difficult to estimate the effect of GATT on dyads where one country occupies a marginal position in the world trade network. This is because the probability to participate in the GATT is much higher for more central countries. Unmatched control dyads show also much higher values of the local clustering indicator. Finally, also global cluster membership prevalences differ quite substantially between dyads on and off support. This may indicate that these clusters are able to capture some unobserved characteristics that are also related to the propensity to participate in the GATT.

Finally, Table 4 compares the estimates of ATE obtained using the three propensity score matching procedures. First of all, we notice that ATE is consistently positive and statistically significant in all three propensity score matching approaches compared. However, the estimated average causal effect of GATT is substantially higher when network variables are ignored. According to the first propensity score approach the ATE is estimated to be 0.42, which approximately amounts to say that GATT increases bilateral trade by 52%. Adjusting for all network variables gives a lower average effect of 0.30, corresponding to a 35% increase in bilateral trade. The higher estimate we obtain ignoring network variables may suffer from omitted variable bias. Indeed, we find that there exist systematic differences in countries' network characteristics between the treatment group and the control group, which are still present even after having controlled for dyads' background characteristics, $\boldsymbol{X}$. These differences may lead to biased estimates of the effect of GATT membership, because network characteristics (i.e., local and global centrality and clustering) are correlated with both bilateral trade and GATT membership.

One possible explanation of the relevance of network characteristics is that they capture the relevance of market power in world trade flows (Eicher and Henn, 2011). The heterogeneity in the level of centrality and clustering of individual countries, resulting in a core-periphery structure, one one hand, and the higher propensity of central countries to be GATT members, on the other hand, indicate that countries that are well connected and that are linked to well connected countries tend to participate more to multilateral trade liberalization because of their relative advantage on world markets. Moreover, market power is a variable notoriously difficult to measure, even more at a very aggregated level. The position of a country in the structure of international trade relations can reveal the comprehensive as well as latent dimension of market power.

Table 3: Means/proportions per cent of pre-treatment variables by dyads on and off common support and treatment status with three different specifications of the propensity score

	The propensity score model includes:											
	$\boldsymbol{X}$				$\boldsymbol{X}$ *and* $\tilde{C}_D$				$\boldsymbol{X}$ *and* $\boldsymbol{N}$			
	Group				*Group*				*Group*			
Variables	*Treated ON*	*Treated OFF*	*Control ON*	*Control OFF*	*Treated ON*	*Treated OFF*	*Control ON*	*Control OFF*	*Treated ON*	*Treated OFF*	*Control ON*	*Control OFF*
Dyad's characteristics												
Currency union (yes/no)	2.4	80.0	0.9	-	2.0	16.7	0.9	0.0	2.2	12.5	0.9	0.0
Logarithm of the distance	8.2	7.0	8.1	-	8.2	8.4	8.1	7.2	8.2	8.6	8.1	7.1
Log of the product of real GDP	48.1	47.0	47.3	-	48.0	49.5	47.3	45.5	48.0	49.5	47.3	45.7
Log of the product of real GDP per capita	15.6	14.8	15.4	-	15.5	15.7	15.4	15.2	15.5	15.8	15.4	15.1
Common language (yes/no)	20.8	40.0	26.4	-	19.8	35.7	25.4	90.9	19.4	37.5	24.2	100.0
Land border (yes/no)	3.1	20.0	4.8	-	3.1	4.8	4.9	0.0	3.2	4.2	4.4	18.2
Number of landlocked	0.1	0.0	0.2	-	0.1	0.0	0.2	0.4	0.1	0.0	0.2	0.3
Number of islands	0.4	0.4	0.2	-	0.3	0.8	0.2	0.2	0.3	0.8	0.2	0.0
Logarithm of the product of land areas	25.8	25.2	25.2	-	25.7	26.6	25.3	24.5	25.7	26.9	25.2	25.1
Common Colonizer post 1945 (yes/no)	5.5	20.0	1.6	-	3.9	26.2	1.6	0.0	4.4	18.8	1.6	0.0
Dyad currently in colonial relationship (yes/no)	1.3	60.0	0.3	-	1.0	11.9	0.3	0.0	0.8	12.5	0.3	0.0
Dyad ever in colonial relationship (yes/no)	3.5	60.0	3.4	-	3.3	11.9	3.4	0.0	3.2	12.5	3.5	0.0
Network measures												
Degree centrality (Country i)	0.9	0.8	0.8	-	0.9	0.9	0.8	0.4	0.9	0.9	0.8	0.4
Degree centrality (Country j)	0.6	0.5	0.6	-	0.6	0.8	0.6	0.6	0.6	0.7	0.6	0.6
Eigen vector centrality (Country i)	0.9	0.8	0.9	-	0.9	0.9	0.9	0.5	0.9	1.0	0.9	0.5
Eigen vector centrality (Country j)	0.7	0.6	0.7	-	0.7	0.8	0.7	0.7	0.7	0.8	0.7	0.7
Local (Country i)	0.7	0.7	0.7	-	0.7	0.7	0.7	1.0	0.7	0.7	0.7	1.0
Local (Country j)	0.8	0.9	0.9	-	0.8	0.8	0.9	0.8	0.8	0.8	0.9	0.8
Cluster membership (Country i): 3 clusters												
Cluster 1	48.7	60.0	38.5	-	48.3	54.8	38.1	63.6	47.9	58.3	37.8	63.6
Cluster 2	21.7	20.0	32.8	-	20.9	31.0	32.3	63.6	23.0	8.3	31.7	68.2
Cluster membership (Country j): 3 clusters												
Cluster 1	14.1	0.0	21.9	-	14.7	4.8	21.9	27.3	14.9	4.2	21.5	36.4
Cluster 2	20.3	20.0	30.2	-	21.5	4.8	30.2	27.3	22.2	0.0	30.1	31.8

Table 4: Estimated ATE and standard error with three different specifications of the propensity score.

The PS model includes	*ATE*	*Standard error*
$\boldsymbol{X}$	0.42	0.11
$\boldsymbol{X}$ and $\tilde{C}_D$	0.28	0.12
$\boldsymbol{X}$ and $\boldsymbol{N}$	0.30	0.13

5 Conclusions

In 1947 the General Agreement on Tariffs and Trade (GATT) was established to regulate international trade among member countries and to reduce tariffs and other trade barriers. It has been long debated whether GATT succeeded in favouring international trade. Recent studies provide mixed results, ranging from the "no impact on trade flows" of the seminal contribution of Rose (2004), to the "strong impact" of Liu (2009), and passing by the heterogeneous effects highlighted by Subramanian and Wei (2007); Gowa and Kim (2005); Eicher and Henn (2011). In this paper we argue that an important confounding factor in the relationship between GATT and bilateral trade has been overlooked in the literature. This relates to the the position of a country in the network of world trade relations, its direct and indirect connections, its tendency to cluster with its neighbours. These network characteristics may convey relevant information on latent characteristics of a country that can contribute explaining its propensity to participate in international agreements. Disregarding network information could seriously bias the estimate of the causal effect of the GATT.

We re-consider the analysis of the effect of the GATT on bilateral trade focussing on comparing bilateral trade of dyads of countries where both participate in the GATT (treated) with bilateral trade of the others dyads (control). We adopt the framework for causal inference known as Rubin Causal Model and show how network information can be incorporated in a propensity score matching analysis of the effect GATT on bilateral trade, under the assumption of strong ignorability given network measures and pre-treatment variables.

Our approach is based on three steps. First, relevant network characteristics have to be summarized using a set of network indicators. For our application we consider a set of four network measures: degree and eigenvector centrality, local clustering indicator, global clustering measures. Second, a propensity score model has to be estimated including network measures. Once an acceptable balance of all covariates and network characteristics has been obtained, outcome data can be analysed.

We show that by adjusting the propensity score for network characteristics, together with more traditional covariates suggested by economic gravity models of bilateral trade, can highlight important differences between treated and control dyads that otherwise could not be detected. In particular, we find that countries in control dyads tend to have less connections, to be more peripheral to the network of international trade and tend to have a higher propensity to form local clusters as compared to countries in treated dyads. Insisting on comparing treated and control dyads ignoring these difference with respect to network characteristics, as traditionally done, may introduce a substantial bias in estimating the effect of GATT. We find some evidence in this respect.

We compare estimates obtained using three alternative propensity score models that included: 1) traditional covariates only, $\boldsymbol{X}$; 2) $\boldsymbol{X}$ and the normalized degree centrality of both countries in a dyad; 3) $\boldsymbol{X}$ and all the four network measures considered, $\boldsymbol{N}$. We find that the first approach that ignores network variables gives considerably higher estimates of the effect of GATT on bilateral trade compared to the other two approaches that give similar estimates. We stress that the first propensity score model is not able to obtain an acceptable balance of network characteristics failing to adjust for relevant information required for strong ignorability. Interestingly, we find that in our empirical analysis adjusting only for degree centrality (using the second propensity score model) was sufficient to obtain a good balance also for other network measures, and in particular for eigenvector centrality and local clustering. This results could be highly data dependent but could also be a sign that a limited amount of network statistics is sufficient to capture the structural dimension of the phenomenon under study (Faust, 2007), as also indicated by the very similar estimates of ATE obtained using degree centrality only ($\boldsymbol{X}$ and $\tilde{C}_D$) or all the network variables ($\boldsymbol{X}$ and $\boldsymbol{N}$) in the propensity score. Our network is characterised by a quite high density and, as expected theoretically, the centrality and local clustering measures show rather high correlations. This may explain why balancing degree centrality also helps balancing the other two measures, which is not the case for the global clustering measure.

Several interesting avenue of future research can further develop some of the aspects just exposed in this contribution. The sign of the bias and its invariance along time are worth exploring, to give account of the causal effect of GATT membership for different groups of countries, in different Rounds of negotiations, with a different proportion between treated countries and non-treated ones. Future studies can also use Monte Carlo simulations to investigate the extent of bias introduced when network characteristics are ignored in different network structures.

Simulation studies can also be of help in defining which network measures are sufficient, necessary or better suited to capture the network structure. Here the literature on Exponential Random Graph Models (Robins et al., 2007; Butts, 2009) can act as a stepping stone for the analysis.

Finally, in our application we maintain the SUTVA focusing on the use of network information in propensity score matching. Causal inference studies in the presence of interference are not yet common, especially in observational studies, although it is a cutting-edge research topic that are drawing increasing interest (Hong and Raudenbush, 2006; Sobel, 2006; Rosenbaum, 2007; Kao et al., 2012; Aronow and Samii, 2012; Arpino and Mattei, 2013). A valuable topic for future research is to exploit network information to allow for interference among units in a causal setting.

Table A.1. Percent Absolute Standardized Bias of each variable with three different specifications of the propensity score.

	ASB%	*ASB% after matching*		
	before	*The PS model includes:*		
Variable	*matching*	$\boldsymbol{X}$	$\boldsymbol{X}$ *and* $\tilde{C}_D$	$\boldsymbol{X}$ *and* $\boldsymbol{N}$
Dyad's characteristics				
Currency union (yes/no)	15.5	1.4	0.3	3.4
Logarithm of the distance	18.0	0.2	0.6	0.4
Log of the product of real GDP	39.1	3.0	5.0	1.4
Log of the product of real GDP per capita	16.4	1.3	4.8	3.4
Common language (yes/no)	12.7	1.7	0.8	2.3
Land border (yes/no)	8.0	0.7	0.5	1.9
Number of landlocked	28.8	0.7	3.4	6.6
Number of islands	24.9	1.2	3.7	3.4
Logarithm of the product of land areas	21.6	0.7	3.3	1.1
Common Colonizer post 1945 (yes/no)	21.8	2.9	2.2	2.4
Dyad currently in colonial relationship (yes/no)	15.3	4.1	1.8	0.7
Dyad ever in colonial relationship (yes/no)	3.1	1.0	1.8	3.0
Network measures				
Degree centrality (Country i)	40.3	25.0	3.8	3.1
Degree centrality (Country j)	34.0	17.5	2.7	3.6
Eigen vector centrality (Country i)	39.4	24.3	3.1	2.7
Eigen vector centrality (Country j)	32.5	14.8	0.0	-4.2
Local (Country i)	41.7	25.5	4.4	3.4
Local (Country j)	37.1	22.0	8.2	3.1
Cluster membership (Country i): 3 clusters				
Cluster 1	20.9	9.0	9.1	4.2
Cluster 2	25.0	21.8	16.4	5.3
Cluster membership (Country j): 3 clusters				
Cluster 1	21.0	12.5	6.5	0.3
Cluster 2	22.9	30.0	22.8	1.1

References

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics 3*(1), 133–160.

Aronow, P. M. and C. Samii (2012). Estimating average causal effects under general interference. In *Summer Meeting of the Society for Political Methodology, University of North Carolina, Chapel Hill, July*, pp. 19–21. Citeseer.

Arpino, B. and A. Mattei (2013). Assessing the impact of financial aids to firms: Causal inference in the presence of interference. In *MPRA Working Paper*, Number 51795.

Bagwell, K. and R. W. Staiger (2004). *The Economics of the World Trading System.* MIT Press.

Baier, S. L. and J. H. Bergstrand (2009). Estimating the effects of free trade agreements on international trade flows using matching econometrics. *Journal of International Economics 77*(1), 63–76.

Baldwin, R. E. and A. J. Venables (1995). Regional economic integration. In G. M. Grossman and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 3, Chapter 31, pp. 1597–1644.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*, 41–55.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science 325*(5939), 414–416.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22*(1), 31–72.

Christakis, N. A., J. H. Fowler, G. W. Imbens, and K. Kalyanaraman (2010). An empirical model for strategic network formation. Technical Report 16039, National Bureau of Economic Research.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

De Benedictis, L., S. Nenci, G. Santoni, L. Tajoli, and C. Vicarelli (2014). Network analysis of world trade using the BACI-CEPII dataset. *Global Economy Journal 14*(3-4), 287–343.

De Benedictis, L. and D. Taglioni (2011). The gravity model in international trade. In L. De Benedictis and L. Salvatici (Eds.), *The Trade Impact of European Union Preferential Policies*, Chapter 4, pp. 55–89. Springer.

De Benedictis, L. and L. Tajoli (2011). The world trade network. *The World Economy 34*(8), 1417–1454.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Di Nardo, J. and J. L. Tobias (2001). Nonparametric density and regression estimation. *Journal of Economic Perspectives 15*(4), 11–28.

Doreian, P. (2001). Causality in social network analysis. *Sociological Methods & Research 30*(1), 81–114.

Eicher, T. S. and C. Henn (2011). In search of WTO trade effects: Preferential trade agreements promote trade strongly, but unevenly. *Journal of International Economics 83*(2), 137–153.

Fagiolo, G. (2015). The international trade network: Empirics and modeling. In J. Nicoll Victor, M. Lubell, and A. H. Montgomery (Eds.), *Oxford Handbook of Political Networks*, Chapter 29. Oxford University Press.

Faust, K. (2007). Very local structure in social networks. *Sociological Methodology 37*(1), 209–256.

Goldstein, J. L., D. Rivers, and M. Tomz (2007). Institutions in international relations: Understanding the effects of the GATT and the WTO on world trade. *International Organization 61*(1), 37–67.

Gowa, J. and S. Y. Kim (2005). An exclusive Country Club: the effects of the GATT on trade, 1950–94. *World Politics 57*(4), 453–478.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. *Handbook of International Economics 4*, 131–195.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies 65*(2), 261–294.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies 64*(4), 605–654.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–970.

Hong, G. and S. W. Raudenbush (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association 101*, 901–910.

Imai, K. and I. Kim (2015). On the use of linear fixed effects regression estimators for causal inference. *Working Paper*.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, NY, USA: Cambridge University Press.

Irwin, D. A. (1996). *Against the Tide: An Intellectual History of Free Trade.* Princeton University Press.

Irwin, D. A., P. C. Mavroidis, and A. O. Sykes (2008). *The Genesis of the GATT*. Cambridge, UK: Cambridge University Press.

Jackson, M. O., B. W. Rogers, and Y. Zenou (2015). The economic consequences of social network structure. *Working paper available at SSRN*.

Kao, E., R. Toulis, E. Airoldi, and B. D. Rubin (2012). Causal estimation of peer influence effects. Paper presented at the *NIPS 2012 Workshop 'Social Network and Social Media Analysis: Methods, Models and Applications'*.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models.* Springer Science & Business Media.

Lee, L. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*(2), 333–374.

Liu, X. (2009). GATT/WTO promotes trade strongly: Sample selection and model specification. *Review of International Economics 17*(3), 428–446.

Lusher, D., J. Koskinen, and G. Robins (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications.* Cambridge University Press.

Maggi, G. (2014). International trade agreements. In E. Helpman, K. Rogoff, and G. Gopinath (Eds.), *Handbook of International Economics*, Volume 4, pp. 317–390.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*(23), 8577–8582.

Normand, S.-L. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of clinical epidemiology 54*(4), 387–398.

O'Rourke, K. H. and J. G. Williamson (2001). *Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy*. Mit Press.

Reichardt, J. and S. Bornholdt (2006). Statistical mechanics of community detection. *Physical Review E 74*(1), 1–16.

Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison (2007). Recent developments in exponential random graph (p*) models for social networks. *Social networks 29*(2), 192–215.

Rose, A. K. (2004). Do we really know that the WTO increases trade? *The American Economic Review 94*(1), 98–114.

Rose, A. K. (2007). Do we really know that the WTO increases trade? reply. *The American Economic Review 97*(5), 2019–2025.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*, 191–200.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician 3*, 33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psycology 66*, 688–701.

Rubin, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics 2*, 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects. *The Annals of Statistics 6*, 34–58.

Rubin, D. B. (1990). Comment: "Neyman (1923) and causal inference in experiments and observational". *Statistical Science 5*, 472–480.

Rubin, D. B. (2008). For objective causal inference design trumps analysis. *The Annals of Applied Statistics 2*(3), 808–840.

Smith, J. (2000). A critical survey of empirical methods for evaluating active labor market policies. *Swiss Journal of Economics and Statistics (SJES) 136*(III), 247–268.

Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics 125*(1), 305–353.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association 101*, 1398–1407.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1), 1–21.

Subramanian, A. and S. Wei (2007). The WTO promotes trade, strongly but unevenly. *Educational Researcher 72*(1), 151–175.

Tomz, M., J. L. Goldstein, and D. Rivers (2007). Do we really know that the WTO increases trade? Comment. *The American Economic Review*, 2005–2018.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*(6684), 440–442.