

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

Bruno Arpino

*Department of Political and Social Sciences, Universitat Pompeu Fabra
Edifici Jaume I, Ramon Trias Fargas, 25-27, E-08005 Barcelona, Spain*

Luca De Benedictis

*Rossi-Doria Center and Department of Economics and Law, University of Macerata
Via Crescimbeni 20, 62100 Macerata, Italy*

Alessandra Mattei

*Department of Statistics, Computer Science, Applications, University of Florence
Viale Morgagni 59, 50134 Florence, Italy*

Abstract. Motivated by the evaluation of the causal effect of the General Agreement on Tariffs and Trade on bilateral international trade flows, we investigate the role of network structure in propensity score matching under the assumption of strong ignorability. We study the sensitivity 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 highly sensitive to the presence of node-level network statistics in the set of confounders. Therefore, we argue that estimates may suffer from omitted variable bias when the relational dimension of units is 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 (Irwin et al., 2008). It was signed in 1947 by 23 nations. Over the subsequent years, eight 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 included as a treaty on trade in goods among the legal pillars of the newly established 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.

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; Tomz et al., 2007; Subramanian and Wei, 2007; Eicher and Henn, 2011). The empirical approach usually adopted to evaluate the effect of GATT/WTO membership on bilateral trade is based on cross-sectional or longitudinal linear fixed effects regression methods, including either unit specific effects or unit and time specific effects (e.g., Rose, 2004; Tomz et al., 2007). This approach presents some shortcomings: the effects of interest are generally not stable over time (Baier and Bergstrand, 2009) and generally lack of a causal interpretation unless specific assumptions are invoked. In this paper we adopt an alternative approach, framing causal inference in the context of the potential outcome approach, usually referred to as Rubin's Causal Model (RCM, see Imbens and Rubin, 2015, for a comprehensive overview of the RCM). In the potential outcome approach, observational studies have to be carefully planned to approximate randomized experiments to obtaining "objective" causal inferences (e.g., Rubin, 2008). Pre-treatment information is crucial to properly design observational studies. In some cases this information is relevant but it is unobservable or difficult to obtain and it is often approximated, at best, or fully ignored, at worst. A common case of unobserved information is the one associated to social network structure. If the network structure is of relevance, the position of a unit (e.g., 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, the development and consolidation of statistical techniques 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) and the second one on peer effects (Manski, 1993; 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 structures in causal inference, and the issue of causality is still a very much debated one in network analysis (Doreian, 2001). In this paper we make a step forward in the direction of accounting for network structures in observational studies of causal inference.

Under the RCM, causal inference involves 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. With respect to the main body of the literature, that focus on the average effect of GATT in a time span half century long, we focus on one specific instant of time, evaluating the effect of GATT membership in 1954 on bilateral trade in 1955.

The second part 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. In the literature, this part is usually referred to as ‘the design’ of studies for causal effects and involves “all contemplating, collecting, organizing, and analysing of data that takes place prior to seeing any outcome data” (Rubin, 2008, page 810). The design of observational studies aims at approximating randomized experiments as closely as possible. To this end we invoke the assumption of strong ignorability (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 of including pre-treatment information on the network structure of the data in the set of variables conditional on which strong ignorability is reasonably assumed. Under strong ignorability, 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 estimate 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 briefly review the literature on GATT/WTO, we describe the data, and we define the world trade network and the statistics used to synthetically represent the structure of world trade. In Section 3 we describe the methodological framework and in Section 4, we present and discuss the results. Section 5 concludes.

2. International Trade and Trade Policy: Issue and data

In spite of international trade and commercial policy being one of the oldest topics of debate in the history of economic thought, the empirical literature on the trade-effect of preferential trade agreements has been dramatically sparse until the publication of the seminal paper by Andrew K. Rose in 2004. Rose (2004) examined the effect of GATT/WTO membership on bilateral trade flows using a longitudinal linear fixed effects regression model. The main result of the analysis was that the GATT had virtually no impact on trade flows between 1947 and 1999.

This was a striking result, given that the GATT was established with 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” (General Agreement on Tariffs and Trade, 1961), and that the Most-favoured-nation principle (GATT, art.1) was guaranteeing that the lowest level of trade barriers was applied among country members. Rose’s thought-provoking result encouraged the research on the topic and produced a series of contributions along the same methodological line (Rose, 2007; Tomz et al., 2007; Subramanian and Wei, 2007; Liu, 2009; Eicher and Henn, 2011, e.g.,) that, as summarized by Maggi (2014), “...overturned or qualified Rose’s results.”

Our analysis stands on the shoulders of these contributions, and also owes some credit to Baier and Bergstrand (2009), who were the first to apply propensity score matching techniques to the analysis of the effect of trade policy on trade flows. Nevertheless neither of these contributions considers the role played by the network structure of trade data.

To be consistent with the original set of information, we use a subset of the dataset built by Rose (2004), which covers bilateral trade flows for 178×177 country pairs, between 1948 and 1999. Specifically, we focus on data from 1954 and 1955. The outcome variable of main interest is a measure of yearly bilateral trade: the average value of bilateral trade between a pair of countries, say 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. Countries that did not report any trade flows were originally excluded from the analysis. The dataset also comprises a wealth of information on background variables, including variables now considered standard in the gravity model literature (e.g., GDP of exporting and importing countries, bilateral distance, contiguity) as well as additional variables that may be critical to reduce the possibility of an omitted variable bias. We provide a complete list of the

variables included in the dataset in the web-based Appendix. Finally, we have information on GATT/WTO membership status of each country. There exist alternative definitions of GATT membership. Rose (2004) focused only on formal membership, which may justify at least partially the unexpected result of no effect he found. We adopt instead the classification by Tomz et al. (2007) who classified as GATT members both formal GATT members and nonmember GATT participants.

2.1. 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 (De Benedictis and Tajoli, 2011). 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.

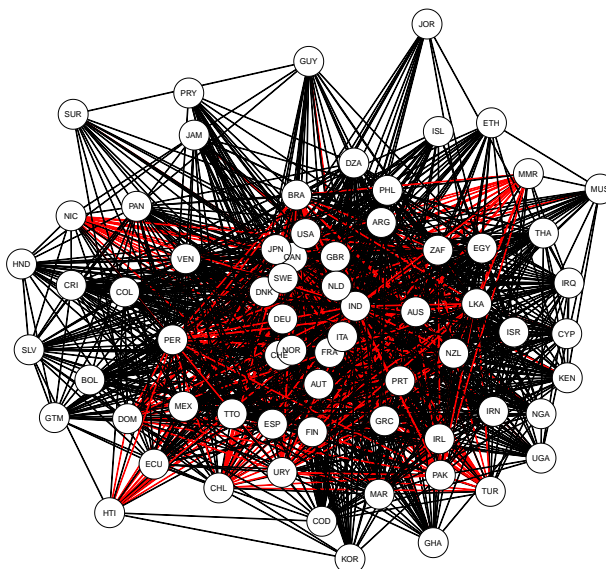
Given the characteristics of the data built by Rose (2004) and Tomz et al. (2007), the world trade network is an undirected graph $G(\mathcal{N}, \mathcal{E})$ that consists of a set of nodes, identifying potential trade partners $\mathcal{N} = \{1, \dots, n\}$, and a list of unordered pairs of nodes, called edges, that correspond to realized trade partnerships: $\mathcal{E} = \{(i, j), (k, l), (j, l), \dots\}$ for i, j, k , and $l \in \mathcal{N}$. The graph G can be also expressed in matrix form by its adjacency matrix $\mathcal{L} = [\mathcal{L}_{ij}]$ where $\mathcal{L}_{ij} = 1$ if $\{i, j\} \in \mathcal{E}$, that is, if countries i and j are trade partners (regardless of the direction of trade flow), and $\mathcal{L}_{ij} = 0$ if $\{i, j\} \notin \mathcal{E}$, that is, if countries i and j are not trading between them.

Figure 1 shows the visualization of G in 1954 (G^{1954}). The red edges in Figure 1 indicate that both countries connected by the 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 $\delta^{1954} = \frac{1319}{0.5(66 \times 65)} = 0.615$. The world trade network, as represented by the data, assumes a core-periphery structure, with some countries taking a central position while others lagging at the boundaries of the network.

In order to quantify the position of countries in the trade network we rely on two standard measures of network centrality (Wasserman and Faust, 1994): the degree centrality, and the eigenvector centrality.

Degree centrality is the simplest measure of the position of a node in a network and measures how a node is connected to its direct neighbours. It is a measure of local centrality. Since the trade network is considered in its binary version, the degree centrality

Figure 1. Trade Partners in 1954 and GATT membership



Note: The figure represents the graph $G(\mathcal{N}, \mathcal{E})$ of trade partners in 1954. 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). Black edges indicates that trade partners were not bot members of the GATT. 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).

measures the centrality of a node by the number of connections the node has. Formally,

$$C_{Di} = \#\{j : \mathcal{L}_{ij} = 1\} = \sum_{j=1, j \neq i}^n \mathcal{L}_{ij}, \quad (1)$$

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

The degree centrality measure, being dependent on the number of the existing nodes in the network, is usually normalized for comparability reasons. We use the maximum degree centrality realized among all the nodes h , $\max_h \#\{h : \mathcal{L}_{ih} = 1\}$, as a normalizing factor. Formally,

$$\tilde{C}_{Di} = \frac{\sum_{j=1, j \neq i}^n \mathcal{L}_{ij}}{\max_h \sum_{j=1, j \neq h}^n \mathcal{L}_{ih}}. \quad (2)$$

It follows that this indicator ranges from 0 to 1; the more the degree centrality of a country is close to 1, the more that country is connected in the network. For example, in 1954 Chile (CHL) has $\tilde{C}_{DCHL} = 0.661$; while Great Britain (GBR) has $\tilde{C}_{DGBR} = 1$.

The eigenvector centrality is instead an index based on direct and indirect neighbours' characteristics, taking into account how important, central, influential or tightly clustered the node's neighbours are. The eigenvector centrality is therefore a measure of global centrality, in which the position of each node's neighbours, and the direct and indirect links of a node, all contribute to 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.

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:

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

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

$$(I - \mathcal{L}) \vec{C}_E = 0, \quad (4)$$

where I is a $n \times n$ identity matrix, and \vec{C}_E is the $n \times 1$ vector of countries' eigenvector centralities (Equation 4 is the trade-matrix characteristic equation for an eigenvalue $\lambda=1$). 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, 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:

$$\tilde{C}_{Ei} = \frac{C_{Ei}}{\max_h C_{Eh}}. \quad (5)$$

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

In general, a high level of \tilde{C}_{Ei} corresponds to countries belonging to large and cohesive (high-density) sub-networks, or, in other terms, countries with a high value of \tilde{C}_{Ei} 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_{Di} neighbours, the number of possible edges that could exist among the nodes within the neighbourhood is $C_{Di}(C_{Di} - 1)/2$. Thus, the

local clustering coefficient for country i can be defined as:

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

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 is directly linked to any other node that is connected to i .

The measure of global clustering that we consider, denoted by C_{Ci} , is generated using a “community detection algorithm” (Newman, 2006) and in particular the Spin Glass algorithm by 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. Therefore, in our propensity score models we include two dummy variables indicating whether a country belongs or not to the first or second class, respectively (C_{C1i} and C_{C2i}).

All these measures, \tilde{C}_{Di} , \tilde{C}_{Ei} , C_{Ti} , C_{C1i} and C_{C2i} , 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 this section we introduce and define the basic concepts for causal inference within the RCM: units, treatments and potential outcomes.

In our study units are country-dyads (i, j) in $t = 1954$. The treatment variable of interest is belonging to the GATT in 1954, either as a formal member or as non-member participants (Tomz et al., 2007). The choice of the specific year of analysis is carefully made in order to guarantee a good proportion between GATT members and non-GATT members in the world trade network, and a stable number of GATT members in a time-interval around t . 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 member 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 comprises either only one GATT member or no GATT member, and so $Y_{ij}(0)$ can be unstable. We investigate the sensitivity of our results with respect to the no hidden versions of treatments component of SUTVA by also conducting the analysis using three alternative definitions of treatment: (1) $T_{ij} = 1$ if both countries in dyad (i, j) are GATT members, and $T_{ij} = 0$ if only one country in dyad (i, j) is a GATT member; (2) $T_{ij} = 1$ if both countries in dyad (i, j) are GATT members, and $T_{ij} = 0$ if no country in dyad (i, j) is a GATT member; and (3) $T_{ij} = 1$ if only one country in dyad (i, j) is a GATT member, and $T_{ij} = 0$ if no country in dyad (i, j) is a GATT member. As an alternative, we could consider a three-level treatment variable. Although our approach could be, at least in principle, easily extended to the case of multivalued treatments, in our study the use of a three-level treatment variable cannot be implemented due to data constraints.

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 the context of trade policy, 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 are, 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:

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

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. 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, \mathbf{X}_{ij} . The vector of the observed covariates, \mathbf{X}_{ij} , includes both dyad-specific information, such as the indicator for common language and the indicator for currency union membership, as well as dyad information based on country-specific characteristics, such as the logarithm of product of the real GDP for country i and country j (see the web-based Appendix for the list of the pre-treatment variables we use in the analysis). Unfortunately, the original data of Rose (2004) and Tomz et al. (2007) does not include country-specific information, but it is worth noting that country-specific characteristics might provide valuable information. 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 the measures introduced in Section 2.2: the normalized degree centrality, \tilde{C}_{Di} , the normalized eigenvector centrality, \tilde{C}_{Ei} , the local clustering coefficient, C_{Ti} and the set of indicators for clusters of nodes derived from community detection algorithms, C_{C1i} and C_{C2i} . Let $\mathbf{N}_i = [\tilde{C}_{Di}, \tilde{C}_{Ei}, C_{Ti}, C_{C1i}, C_{C2i}]$ denote the vector of network measures for country i , $i = 1, \dots, n$.

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 the original definition given by Rosenbaum and Rubin (1983) to the case of an observational study with network data. Formally

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

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

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

Strong ignorability amounts to assuming 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. The overlap assumption guarantees that in large samples we can find treated and control units for all values of the pre-treatment variables and network measures.

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. Alternative adjustment methods can be used, including regression and model-based imputation methods, matching, propensity score methods and combinations of matching and propensity score methods with covariance adjustment. We opt for propensity score methods. 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: $e_{ij} = P(T_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{N}_i, \mathbf{N}_j)$. We view propensity score methods as relatively attractive for two main reasons. First, they do not require building models for the potential outcomes as instead regression and model-based imputation methods do. The second reason is related to the properties of the propensity score (Rosenbaum and Rubin, 1983), 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: $T_{ij} \perp (\mathbf{X}_{ij}, \mathbf{N}_i, \mathbf{N}_j) | e_{ij}$; 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 Assumption 1 holds then $T_{ij} \perp (Y_{ij}(0), Y_{ij}(1)) | e_{ij}$ and $0 < P(T_{ij} = 1 | e_{ij}) < 1$. These properties imply that we can avoid adjusting directly for all covariates, and instead focus on adjusting for differences in the propensity score to create a sample of treated and control observations where pre-treatment background variables and network measures are balanced.

There exist several propensity score methods to create a balanced sample of treated

and control observations, including weighting, where the propensity scores are used as weights, blocking on the propensity score, regression on the propensity score and matching on the propensity score (see e.g., Stuart, 2010; Imbens and Rubin, 2015, for a review). Unfortunately, there is little guidance on how selecting between propensity score methods. We opt for matching on the propensity score, but alternative methods could be used.

As for the general case in 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:

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

where $\boldsymbol{\alpha}$ 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.

Once the propensity score is estimated, several matching methods are available. The most common ones are kernel, nearest neighbour, radius, and caliper (see e.g., Stuart, 2010; Imbens and Rubin, 2015, for a discussion about these methods).

Similarly 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. The absolute standardized bias 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 absolute standardized bias is a measure of covariate balance: a lower absolute standardized bias 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 absolute standardized bias 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 the average treatment effect, 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 the average treatment effect is estimated averaging over all dyads that found at least one matched dyad in the opposite treatment group. Kernel matching consists of matching each treated (control) dyad with all the available control (treated) dyads, with weights depending on the distance in terms of propensity scores between each treated and control dyad. The exact definition of the weights depends on the specific kernel function that it is used, but 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 setting the bandwidth to $h = 0.06$. This type of kernel retains only dyads that do not differ more than h units of propensity score.

Note that, before implementing the matching, we discard dyads with extreme estimated propensity scores to better 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

We present results for three different cases that differ with respect to the set of variables included in the propensity score model: (a) X only, where X is the matrix of the background variables; (b) X plus \tilde{C}_D (degree centrality matrix $\tilde{C}_D = [\tilde{C}_{Di}, \tilde{C}_{Dj}]$); (c) X plus N (where N is a matrix including all the network variables).

In the first case, we only include the background covariates, 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 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 the balance of the N variables to show to what extent the misspecified propensity score reduces imbalance also on the network variables, 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 X variables. In this case we choose the solution that gives the best balance of the X variables and the degree centrality measure, but

Table 1. Mean and median percent Absolute Standardized Bias (ASB %) for X and N separately; and for the joint set of X and N , under three different specifications of the propensity score

		<i>The propensity score model includes:</i>					
		X		X and \tilde{C}_D		X and N	
ASB% of	<i>Sample</i>	<i>Mean</i>	<i>Median</i>	<i>Mean</i>	<i>Median</i>	<i>Mean</i>	<i>Median</i>
		<i>ASB%</i>	<i>ASB%</i>	<i>ASB%</i>	<i>ASB%</i>	<i>ASB%</i>	<i>ASB%</i>
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
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
X and 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

again we show the balance obtained on all the covariates. Under Assumption 1, average treatment effect estimators based on a propensity score that only includes degree centrality as a network measure would still be biased. Nevertheless, 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 to assess the robustness of estimates of the average treatment effect to the inclusion of only one network measure instead of all of them.

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

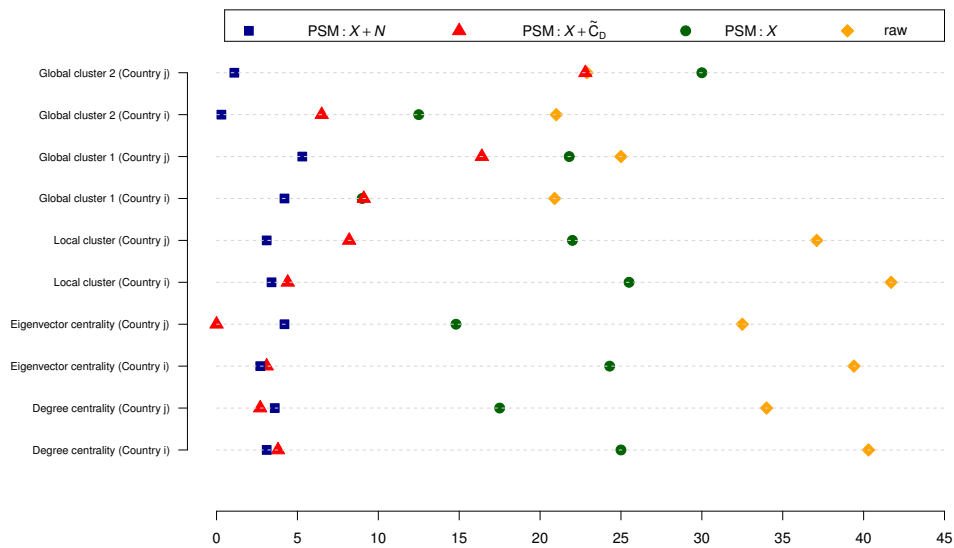
We start presenting the results on the balance of covariates obtained in each of the three cases. For each of the variables we calculate the percent absolute standardized bias before and after matching. The percent absolute standardized bias for each variable is presented in Table A.1 in the web-based Appendix. In Table 1 we summarize this information by presenting the mean and median percent absolute standardized bias calculated on the X variables only, on the N variables only, and on all variables (X and N) under the three specifications of the propensity score.

Let us consider the balance of X variables. Table 1 shows that in the raw dataset

there is a substantial imbalance in the distribution of covariates: the mean and median percent absolute standardized bias are respectively 18.8% and 17.2%. In fact, the absolute standardized bias for several covariates is well above the threshold of 10% (see Table A.1 in the web-based Appendix). The highest absolute standardized bias 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 the average treatment effect, because confounded with the effect of the 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 the absolute standardized bias after matching. All the three propensity score matching procedures succeed in drastically improving the balance of X variables. The mean percent absolute standardized bias for the 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 X variables remains very good in all cases (the maximum percent absolute standardized bias is 6.6% for one covariate in the third case, while it is smaller for all other covariates).

What about the balance of N variables? In Figure 2, we compare the performance of the three matching approaches in terms of reduction in the absolute standardized bias 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 is

Figure 2. Percent Absolute Standardized Bias (ASB%) of network variables with three different specifications of the propensity score model (PSM).



Note: In each row the graph reports the percentage absolute standardized (ASB%) bias for each of the network measures obtained from three different specifications of the propensity score model including: all variables (PSM: X and N); pre-treatment covariates and degree centrality (PSM: X and \tilde{C}_D); and only pre-treatment covariates (PSM: X). In the graph we also report the absolute standardized bias calculated on the raw (unmatched) dataset (raw).

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

<i>Treatment group</i>	<i>The propensity score model includes:</i>						<i>Total</i>
	<i>X</i>		<i>X and \tilde{C}_D</i>		<i>X and N</i>		
	<i>Common support</i>		<i>Common support</i>		<i>Common support</i>		
	<i>Off</i>	<i>On</i>	<i>Off</i>	<i>On</i>	<i>Off</i>	<i>On</i>	
Untreated	0	766	11	755	22	744	766
Treated	5	548	42	511	48	505	553
Total	5	1314	53	1266	70	1249	1319

quite different before matching: the percent absolute standardized bias ranges from 21% to 41.7% (see also Table A.1 in the web-based Appendix). Table 1 indicates that the mean and median percent absolute standardized bias for N are, respectively, 31.5% and 33.3% before matching. As a consequence, when considering the balance of both X and N in Table 1 the mean and median percent absolute standardized bias increase.

The first propensity score matching approach, that ignores network variables, as expected, does not behave well in terms of balance of N . In fact, Figure 2 shows that although there is a reduction in the absolute standardized bias for each of the network variables with respect to the raw dataset, the absolute standardized bias remains quite high (between 20% and 30% for most of these variables). The second propensity score matching approach, which only adds degree centrality, improves considerably the balance with respect to a naive approach that would have not considered the N variables. In this case, in fact, the percent absolute standardized bias of most of the network variables is below 10% with an average of 7.7%. Therefore, in our application 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 absolute standardized bias remains high; only the third propensity score matching approach, where all N variables are used, substantially reduces the imbalance in all the network variables (maximum absolute standardized bias = 5.3%; average absolute standardized bias = 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, discarding unmatched units from the analyses may change the estimand because the unmatched units may hold particular characteristics that make it impossible to compare them with units in the opposite treatment group without relying on model extrapolations (Crump et al., 2009). We investigate the characteristics of the unmatched dyads by calculating 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 and dyads that did not find a match. Results are shown in Table A.2 in the web-based Appendix. From this analysis we get several interesting insights. First, for some covariates, unmatched dyads have quite different average values than matched dyads in the same treatment group. Second, the group of covariates that show quite different average values between matched and unmatched dyads remains stable across the three propensity score models. Third, 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 matched and unmatched dyads. This may indicate that these clusters are able to capture some unobserved characteristics that are also related to the propensity to be a GATT member.

Finally, Table 3 compares the estimates of the average treatment effect, ATE , obtained using the three propensity score matching procedures. First of all, we notice that the average treatment effect is consistently positive and statistically significant in all three propensity score matching approaches compared. However, the estimated average causal

Table 3. Estimated average treatment effect (*ATE*) and standard error with three different specifications of the propensity score model (*PSM*).

<i>The PSM includes</i>	<i>ATE</i>	<i>Standard error</i>
X	0.42	0.11
X and \tilde{C}_D	0.28	0.12
X and N	0.30	0.13

effect of GATT is substantially higher when network variables are ignored. According to the first propensity score approach the average treatment effect 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, 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, on 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 in 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.

In Table 3 the treatment is equal to 1 for dyads where both countries are GATT members in 1954 and 0 otherwise (none or only one is a GATT member in 1954). For illustrative purpose and to gain further understanding on the effect of GATT we also implement our propensity score matching procedure using alternative definitions of the treatment variable. More specifically, we consider three types of comparisons: (1) both countries are in

Table 4. Estimated average treatment effect (ATE) and standard error under alternative definitions of the treatment variable.

<i>Treatment variable</i>	<i>ATE</i>	<i>Standard error</i>
Both in GATT versus one	0.27	0.13
Both in GATT versus none	0.86	0.26
One in GATT versus none	0.37	0.20

the GATT versus only one; (2) both countries are in the GATT versus none; (3) only one country is in the GATT versus none. For each comparison, the propensity score matching procedure outlined above is repeated including all covariates and network variables and excluding one sub-sample in turn (dyads with no, one, and two GATT participants, respectively). Results, reported in Table 4, confirm that GATT participation has a positive effect on bilateral trade and this is true irrespective of the type of comparisons we consider. Nevertheless, as expected, the effect is the strongest when comparing dyads in which both countries participate in the GATT with dyads in which none of the country is a GATT participant ($ATE = 0.86$). The effect of GATT participation when comparing dyads with 2 versus 1 GATT participant is considerably lower ($ATE = 0.27$) and similar to the effect reported in the last row of Table 3 (comparing dyads with two participants versus all the others, $ATE = 0.30$). These results are in line with those found by Tomz et al. (2007) for the same year we analyse (1955; see the second row of Table 5 in their paper). The fact that the (weighted) average of the first two average treatment effects in Table 4 is much more similar to the effect we obtain comparing dyads in which both countries participate in the GATT with dyads in which only one country is in the GATT is probably due to the fact that the size of the group of dyads with no GATT participants is considerably lower than that of the other two groups (about 10%).

5. Conclusions

In the literature there exists a long debate on 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 Eicher and Henn (2011) among others. 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 position of a country in the network of world trade relations, its direct and indirect

connections, and 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 GATT membership.

To shed light on this issue, we re-consider the analysis of the effect of the GATT on bilateral trade focussing on comparing bilateral trade of dyads of countries that both participate in the GATT (treated) with bilateral trade of the others dyads (control). We adopt the potential outcome approach and investigate the role of network information in drawing causal inference under the assumption of strong ignorability. For our application we consider a set of four network measures: degree centrality, eigenvector centrality, a local clustering indicator, and a global clustering measure. We show that network characteristics, together with more traditional covariates suggested by economic gravity models of bilateral trade, can contain crucial information and highlight important differences between treated and control dyads that could not be detected otherwise. 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. Therefore, 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 get some evidence on the role of network information by comparing estimates of the average treatment effects adjusting for three alternative sets of background variables: (a) traditional covariates only, X ; (b) X and the normalized degree centrality of both countries in a dyad; (c) X and all the four network measures we consider, 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 approach 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 was sufficient to obtain a good balance also for the other network measures, and in particular for eigenvector centrality and local clustering. These 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). It is worth noting that our network is characterised by a quite high density and, as it is expected theoretically, the centrality and local clustering measures show rather high correlations. This may explain why balancing degree centrality also helps balancing

eigenvector centrality and the local clustering measure, although some imbalance remains in the global clustering measure.

Several interesting avenues of future research can further develop some of the aspects just exposed in this contribution. It is worth to devote more research on the sign of the bias and its invariance along time, and on the heterogeneity of the causal effect of GATT membership for different groups of countries, in different Rounds of negotiations, and with different proportions between treated and non-treated dyads. 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 help defining which network measures are sufficient, necessary or better suited to capture the network structure. Here the literature on Exponential Random Graph Models (e.g., 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 (e.g., Kao et al., 2012; Arpino and Mattei, 2016). A valuable topic for future research is to exploit network information to allow for interference among units in a causal setting.

References

- Arpino, B. and A. Mattei (2016). Assessing the impact of financial aids to firms: Causal inference in the presence of interference. *The Annals of Applied Statistics*, Forthcoming.
- 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.
- 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.
- 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. and L. Tajoli (2011). The world trade network. *The World Economy* 34(8), 1417–1454.
- 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.
- Faust, K. (2007). Very local structure in social networks. *Sociological Methodology* 37(1), 209–256.
- General Agreement on Tariffs and Trade (1961). *The Activities of GATT 1960/61*. Geneva, Switzerland: GATT.
- Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica* 76(3), 643–660.
- 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., 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’*.
- Liu, X. (2009). GATT/WTO promotes trade strongly: Sample selection and model specification. *Review of International Economics* 17(3), 428–446.
- 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.
- 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. 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. (1990). Comment: “Neyman (1923) and causal inference in experiments and observational studies”. *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.
- 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.

Web-based Appendix:

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

Bruno Arpino

*Department of Political and Social Sciences; Universitat Pompeu Fabra
Edifici Jaume I, Ramon Trias Fargas, 25-27, E-08005 Barcelona, Spain*

Luca De Benedictis

*Rossi-Doria Center and Department of Economics and Law; University of Macerata
Via Crescimbeni 20, 62100 Macerata, Italy*

Alessandra Mattei

*Department of Statistics, Computer Science, Applications; University of Florence
Viale Morgagni 59, 50134 Florence, Italy*

Background variables and additional tables

Complete description of the data and of the data sources is in Rose (2004) and Tomz et al. (2007).

For each dyad (i, j) we have information on the following variables:

- the logarithm of the product of the real gross domestic product (GDP) of country i and country j ;
- the log product of real GDP of country i and country j per capita;
- the logarithm of the bilateral distance between i and j , measured using the great-circle distance formula;
- a dummy variable indicating if country i and country j are sharing a land border;
- 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 one 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 .

The last three variables are omitted in our analysis because of lack of variation in 1954.

References

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

Table A1. Percent Absolute Standardized Bias (ASB%) of each variable with three different specifications of the propensity score

<i>Variable</i>	<i>ASB%</i>	<i>ASB% after matching</i>		
	before matching	<i>The PS model includes:</i>		
		<i>X</i>	<i>X and \tilde{C}_D</i>	<i>X and N</i>
<i>Dyad's characteristics</i>				
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
Logarithm of the distance	18.0	0.2	0.6	0.4
Land border (yes/no)	8.0	0.7	0.5	1.9
Logarithm of the product of land areas	21.6	0.7	3.3	1.1
Number of islands (0,1,2)	24.9	1.2	3.7	3.4
Number of landlocked (0,1,2)	28.8	0.7	3.4	6.6
Common language (yes/no)	12.7	1.7	0.8	2.3
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
Currency union (yes/no)	15.5	1.4	0.3	3.4
<i>Network measures</i>				
Degree centrality (Country <i>i</i>)	40.3	25.0	3.8	3.1
Degree centrality (Country <i>j</i>)	34.0	17.5	2.7	3.6
Eigenvector centrality (Country <i>i</i>)	39.4	24.3	3.1	2.7
Eigenvector centrality (Country <i>j</i>)	32.5	14.8	0.0	-4.2
Local cluster (Country <i>i</i>)	41.7	25.5	4.4	3.4
Local cluster (Country <i>j</i>)	37.1	22.0	8.2	3.1
Global cluster (Country <i>i</i>): 3 clusters				
Cluster 1	20.9	9.0	9.1	4.2
Cluster 2	25.0	21.8	16.4	5.3
Global cluster (Country <i>j</i>): 3 clusters				
Cluster 1	21.0	12.5	6.5	0.3
Cluster 2	22.9	30.0	22.8	1.1

Table A2. Means/proportions per cent of pre-treatment variables by dyads matched and unmatched and treatment status with three different specifications of the propensity score model

<i>The propensity score model includes:</i>												
<i>Variables</i>	<i>X</i>				<i>X and \tilde{C}_D</i>				<i>X and N</i>			
	<i>Group</i>				<i>Group</i>				<i>Group</i>			
	<i>Treated unmatched</i>	<i>Control matched</i>	<i>Control unmatched</i>		<i>Treated unmatched</i>	<i>Control matched</i>	<i>Control unmatched</i>		<i>Treated unmatched</i>	<i>Control matched</i>	<i>Control unmatched</i>	
<i>Dyad's characteristics</i>												
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
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
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
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
Number of islands (0,1,2)	0.4	0.4	0.2	-	0.3	0.8	0.2	0.2	0.3	0.8	0.2	0.0
Number of landlocked (0,1,2)	0.1	0.0	0.2	-	0.1	0.0	0.2	0.4	0.1	0.0	0.2	0.3
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
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
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
<i>Network measures</i>												
Degree centrality (Country <i>i</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 <i>j</i>)	0.6	0.5	0.6	-	0.6	0.8	0.6	0.6	0.6	0.7	0.6	0.6
Eigenvector centrality (Country <i>i</i>)	0.9	0.8	0.9	-	0.9	0.9	0.9	0.5	0.9	1.0	0.9	0.5
Eigenvector centrality (Country <i>j</i>)	0.7	0.6	0.7	-	0.7	0.8	0.7	0.7	0.7	0.8	0.7	0.7
Local cluster (Country <i>i</i>)	0.7	0.7	0.7	-	0.7	0.7	0.7	1.0	0.7	0.7	0.7	1.0
Local cluster (Country <i>j</i>)	0.8	0.9	0.9	-	0.8	0.8	0.9	0.8	0.8	0.8	0.9	0.8
Global cluster (Country <i>i</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
Global cluster (Country <i>j</i>): 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