Journal of Data Science 21 (3), 578–598
July 2023

DOI: 10.6339/22-JDS1064
Data Science in Action

Optimal Physician Shared-Patient Networks and the Diffusion of Medical Technologies

A. James O'Malley[1,*], Xin Ran[2], Chuankai An[3], and Daniel N. Rockmore[4]

[1] *Department of Biomedical Data Science and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA*

[2] *Department of Biomedical Data Science, The Dartmouth Institute for Health Policy and Clinical Practice, and the Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA*

[3] *Research Institute of China Investment Corporation, Beijing, 100010, China*

[4] *Department of Mathematics and Department of Computer Science, Hanover, NH 03755, USA, The Santa Fe Institute, Santa Fe, NM 87501 USA*

Abstract

Social network analysis has created a productive framework for the analysis of the histories of patient-physician interactions and physician collaboration. Notable is the construction of networks based on the data of "referral paths" – sequences of patient-specific temporally linked physician visits – in this case, culled from a large set of Medicare claims data in the United States. Network constructions depend on a range of choices regarding the underlying data. In this paper we introduce the use of a five-factor experiment that produces 80 distinct projections of the bipartite patient-physician mixing matrix to a unipartite physician network derived from the referral path data, which is further analyzed at the level of the 2,219 hospitals in the final analytic sample. We summarize the networks of physicians within a given hospital using a range of directed and undirected network features (quantities that summarize structural properties of the network such as its size, density, and reciprocity). The different projections and their underlying factors are evaluated in terms of the heterogeneity of the network features across the hospitals. We also evaluate the projections relative to their ability to improve the predictive accuracy of a model estimating a hospital's adoption of implantable cardiac defibrillators, a novel cardiac intervention. Because it optimizes the knowledge learned about the overall and interactive effects of the factors, we anticipate that the factorial design setting for network analysis may be useful more generally as a methodological advance in network analysis.

Keywords *bipartite network; directional information; factorial design; implantable cardiac defibrillators; optimal bipartite-unipartite projection; shared-patient physician network*

1 Introduction

Networks encoding the relationships between physicians and patients, and also between physicians (through shared patients) increasingly are used for research in medicine and health care (An et al., 2018b,a, 2019; Barnett et al., 2011). Patient-physician interactions naturally give rise to a *bipartite network* connecting physicians (on one "side") to the patients that they treat (on the other "side") (see Figure 1 (L)). Such a network is clearly a very lossy data construct.

[*] Corresponding author. Email: James.OMalley@Dartmouth.edu.

Received March 22, 2022; Accepted August 15, 2022

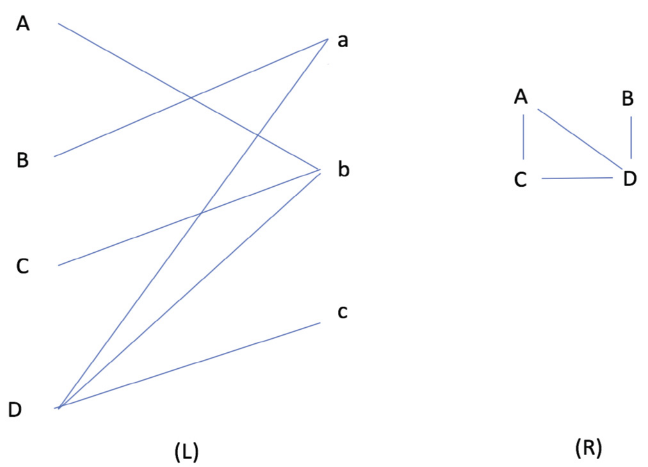

Figure 1: Example Bipartite network and its projection. (L) A toy physician-patient network in which capital letters represent physicians and small letters represent patients with an edge between them encoding a physician-patient relationship. (R) The “single-mode” projection that produces the physician network, connecting physicians who share at least one patient.

Recently, it has been found that the sequences of (presumably linked) patient-physician visits (herein called “referral paths”) may be a more useful encoding of both patient interactions with the health care system as well as a data source for understanding inter-physician interactions. Referral paths can even be used as the basic data of a useful health care-related network construct (An et al., 2018a, 2019).

In this paper we present a way to integrate the more fine-grained information of referral paths with the more traditional bipartite physician-patient network through the network technique of *mode projection* (executed on the physician-patient network). A bipartite network on sets S and T has only edges between nodes in S and nodes in T and not within the sets. “Mode projection” onto S is any form of constructing an individual network on S according to some rule (similarly for T). The simplest form of mode projection links nodes within S if they are each linked to the same node in T. Standard mode projection onto the physicians creates a physician network (a “shared-patient network”) by connecting physicians if they share a patient. But there are other ways to connect physicians that both use and reflect additional information – when available. For example, the threshold used to define binary edges (e.g., only connect two physicians if the number of shared patients exceeds some threshold), has been considered (Farine and Whitehead, 2015). In addition, the sensitivity and specificity of different rules for building shared-patient networks has been estimated (Barnett et al., 2011). Prior work has also considered whether restricting patient visits to physicians visited during the same episode of care leads to networks with more desirable properties (Onnela et al., 2018).

The order in which patients visit physicians – the information that underlies referral paths – appears to largely have been ignored in shared-patient networks. Motivated by the problem of estimating the relationship of hospital networks to their adoption of medical technologies, which we conjecture is informed by features of directed networks, we are interested in whether analyses may be enhanced by exploiting more fully the information in referral paths. In particular, we seek the mappings from the longitudinal bipartite patient-physician referral path space to a unipartite physician network that allows directed network features (e.g., reciprocity, directional assortativity, transitivity) to be formed, optimally discriminates the networks, and improves the predictive accuracy of whether a hospital adopts the technology to perform implantable cardiac defibrillator (ICD) procedures, a cardiac surgery intervention. We will demonstrate the utility of directional information and retaining other forms of information in referral paths using the 8 directed and 8 undirected network features in Table 1 to summarize and compare the topology of each hospital network.

We also regress the outcome in our motivating application – whether a hospital has adopted the technology to perform ICD procedures – on the network features evaluated under each mode projection method. Therefore, we look at the interesting question of technology adoption and diffusion, a subject where network effects have – in other domains – already been shown to be useful (Ryan and Cross, 1943; Griliches, 1957; Coleman et al., 1959; Valente, 1995; Skinner and Staiger, 2005; Rogers, 2010; Iyengar et al., 2011), from the perspective of whether the information in the directionality of referral paths enhances predictive accuracy and the extent to which predictive accuracy is sensitive to which projection method is used. To date the literature only contains examples of sensitivity analyses that evaluate whether study conclusions change over a range of thresholds for reducing a weighted network to a binary network (Schwarz and McGonigle, 2011). Our assessment of sensitivity to the method of forming the network is multi-dimensional and thus much more comprehensive.

Although the motivation for this paper stems from the particular case of patient-physician referral path data, the problem of interest generalizes to any situation in which a directed network is to be constructed from a data set of the sequence of encounters of an actor of one type with actors of another type. Although there is literature on bipartite directed graphs, such as that on Petri nets (Liang et al., 2020), in general surprisingly little has been published about the retention of directional information through the projection to a one-mode network. This is despite the common occurrence of bipartite network data across a wide array of fields (Pavlopoulos et al., 2018), including situations where the bipartite network may naturally be directed or could be represented as such (e.g., authorship order or importance in a collaboration network, timing of queries in internet-based communication networks) but for convenience a projection is often made to an undirected network. Networks with asymmetric weights have been obtained as a consequence of a node's degree or strength in the two-mode network normalizing their edge-weights or use of a relative threshold (e.g., to retain the top 10% of each nodes edges based on degree) to yield binary directed edges (Newman, 2001; Onnela et al., 2018) as opposed to time-ordered paths in the two-mode network.

Building networks from data (even data that is a priori network data) can be challenging due to the many choices available to a network scientist. For example, directed networks are often converted into undirected networks (and even in this conversion there are choices) in order to evaluate network features developed for the latter that don't adapt easily to the former. For network visualization, dense directed networks are often sparsified and made undirected in order to de-clutter presentation. Decisions can be more principled (see e.g., Foti et al. (2011)) and in the setting in which the choices affect some kind of integrative objective function, optimization techniques can be used. Herein we introduce a framework for parameter setting – as well as a principled approach (a designed experiment) to articulating the projection-parameters – that focuses on *discriminatory power* of projection-parameter settings, characterized by the extent to which a vector of network features that summarize the network varies between the resulting networks. Because discriminatory power does not depend on any outcomes (e.g., ICD status), optimizing the network design with respect to it offers protection from cherry-picking a projection method in order to obtain the best result; e.g., maximizing the effect of a network-based predictor on a specific outcome. A contribution of this paper besides the retention of directional information from referral paths is the development of both an objective function quantifying discriminatory power and methodology (outlined in the next paragraph) for efficiently searching for the optimal approach to forming networks via bipartite projections that maximize discriminatory power (and thus avoid being guided by the outcomes the networks will be used to predict and suffering a form of overfitting when selecting the optimal projection).

Given a measurable objective like optimizing discriminatory power across the shared-patient physician networks of a sample of hospitals, we can frame the statistical investigation as an optimization problem. The solutions are intractable in closed form given the enormous search space over the numerous ways of parsing a referral pathway. However, by specifying a set of projection-parameters (or "factors" in design-of-experiments terminology) with fixed (categorical) levels, the computational challenge is reduced to emulating an experiment with a factorial design (a description of factorial designs and experimental designs in general is contained in Montgomery (2017)). Factorial designs systematically vary the levels of the factors across the observations so that every combination is tested. This allows for estimation of the complete set of interactions between the factors using the fewest observations and more generally allows the effect of each factor or combination of factors (interaction) to be estimated with maximum precision.

The structure of the remainder of the paper is as follows. In Section 2 we describe the factors over which we evaluate optimal projections and in Section 3 develop the criteria for evaluating the performance of the projections and the effects of the factors underlying them in terms of discriminatory power and predictive accuracy. In Section 4 we introduce ICDs and explain why understanding mechanisms governing inter-hospital diffusion of ICD utilization is important. The data used to measure hospital shared-patient physician networks and ICD status and the network features of interest are also described. Results are presented in Section 5 and the paper concludes with a discussion of the findings and potential directions for future work in Section 6.

2 Notation and Experimental Design

In this section we present a general methodology for the construction of a physician network based on referral paths. A referral path is represented as a sequence of visits of a patient with a set of physicians; e.g., $\alpha = ABBAC$ where we denote physicians by alphabetic letters and read referral paths from left-to-right. Thus, a referral path with the continuous subsequence AB indicates that a visit to physician A occurred and was followed by a visit to B within a sufficiently short period of time, $t_{\max}$ – and that there were no intermediate physician visits in between. In the motivating Medicare patient example we use $t_{\max} = 30$ days (see Section 4 for justification). Each patient who visits at least two distinct physicians within the timeframe contributes to the set of referral paths while the physicians they visit on these referral pathways are nodes in the (projected) network.

Because a referral path contains information on the order in which a patient visits physicians, the extent to which one physician refers to another may be quantified. Consider the referral path $\alpha = ABCBBAAABC$. The standard approach (i.e., traditional single mode projection of the bipartite physician-patient network) to forming a shared-patient network would generate from α a simple undirected triad in the physician network of the physicians A, B, and C. Among the information in the referral path that this neglects are the multiple return visits (one from B to B, and several consecutive repeated visits to A). Should the projection have loops? Should the projection have directionality? Specifically, should the projection distinguish

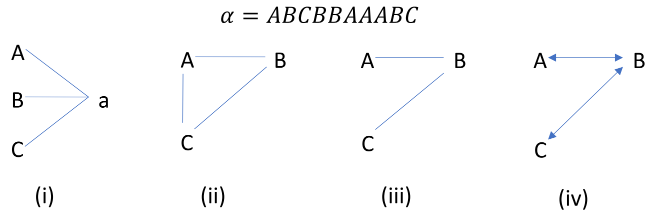

Figure 2: At the top of the figure is an example referral path α for patient "a". The (sub-) patient-physician network is shown in (i). The (sub-) shared-patient network generated by α is shown in (ii), the (undirected) referral relation is encoded in (iii), and directionality is encoded in (iv).

the transitions AB and BA? Other kinds of information involving the order, inter-visit time, total number of visits, and other aspects of the referral path are also elided in the standard projection, potentially a substantial loss of information. Figure 2 gives four different possibilities for projections including one that is directed.

In short, referral paths can form the basis of a variety of ways to measure and articulate the depth of a working relationship or collaboration between physicians and those choices might matter depending on the analytic goal. We seek networks with greater specificity by filtering out uninformative visits while retaining information on directionality. The following considerations, among the many that could be considered, form the basis of our study (keeping in mind that "referral" means the temporal contiguity of physician visits encoded in the data):

- *Continuity* (binary): Is only a direct referral a signal of collaboration or is the simple presence of visits in a given path enough to signal a relationship?
- *Revisit* (binary): Does a "closed loop" of referral indicate a stronger relationship of collaboration and should that contribute to encoding of the relationship?
- *Multiplicity* (five levels): Is the number of referrals in a path important and should that be memorialized in the encoding of an assumed physician-physician relationship?
- *Directed* (binary): Is the directionality of the referral important?
- *Binary* (binary): Is it useful to try to encode the strength of a relationship or is a record of its presence enough? Is there some threshold of interaction that the encoding of collaboration should respect?

The rationale for each of these design-parameters or factors is described by Subsections 2.1 to 2.5 along with their full mathematical specifications. While these considerations are hardly exhaustive, they provide a useful basis to begin our exploration of the relationship between patient referral information, physician collaboration networks, and ultimately, of their utility in helping to understand the structure and efficacy of health care systems. As we use the referral data to build physician collaboration networks in which the five factors identified above are systematically varied across their ranges, we emulate a factorial experiment (see e.g., (Montgomery, 2017, Chapter 5)). All but *Multiplicity* are binary giving a total of $80 = (5 * 2 * 2 * 2 * 2)$ different methods of constructing the physician collaboration network from the patient referral data. In the following subsections we discuss each of these factors by describing the way in which a given factor setting directs the contribution of an individual referral path to the one-mode projected physician network. While it is useful – and intuitive – to think of this visually as the transformation of one network into another, it is perhaps better to re-frame the discussion and presentation in terms of the *adjacency matrix* attached to the projection.

Recall that the *adjacency matrix* of a given network $\mathcal{N}$, is a square matrix $\mathcal{X}(\mathcal{N})$ whose rows and columns are indexed by its vertices. In the case of a physician network, the vertices would be labeled by physicians. In the associated adjacency matrix the entry in row A and column B, $\mathcal{X}(\mathcal{N})_{AB} = \mathcal{X}_{AB}$ represents the weight of the edge from A to B (and equal to 0 when there is no such edge). Unweighted networks have binary adjacency matrices with entries 0 or 1 indicating the absence or presence of an edge. Undirected networks have symmetric adjacency matrices – encoding the lack of direction as edges in both directions (and of equal weights). For a thorough treatment of basic networks and network science see the texts Wasserman and Faust (1994) and Newman (2010).

On the one hand we can think of this process as contributing an edge that joins a pair of physicians A and B (directed or not, thick or thin depending on the weight) or as contributing to the weight recorded in the adjacency matrix. In what follows we'll go back and forth – without apology – between those two descriptions. The adjacency matrices encoding the example networks in Figure 2 (ii) and (iii) (with indices arranged in alphabetical order) are

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Again, it is useful to think of what follows as a set of algorithms that depend on five parameter choices to produce entries for the adjacency matrix of a physician network. We now give some detail on the various parameter spaces.

2.1 Continuity

After visiting physician A, the next physician that the patient visits on referral path α is most likely to have resulted from a referral from physician A, with a decreasing likelihood – although not impossibility – for those physicians visited further along the path. We let *Continuity* equal 1 or 0 depending on whether we do or do not want to enforce the next visit condition for designating a referral. Setting *Continuity* $= 1$ enforces the condition that after a visit to A the immediate next visit must be to B for that segment of α to contribute to the edge from A to B. In contrast, if *Continuity* $= 0$ then any subsequent visit to B qualifies. If *Continuity* $= 1$ then α is decomposed into one or multiple shorter referral paths involving only consecutive visits to A and B. If *Continuity* $= 0$ then α is reduced to a single path by stripping out all physicians not equal to A or B before further processing. For example, $\alpha = ABCA$ is reduced to AB when *Continuity* $= 1$ and to ABA when *Continuity* $= 0$. As a second example, if *Continuity* $= 1$ then $CDABCBAA$ is reduced to AB and BAA (two separate paths) whereas if *Continuity* $= 0$ the AB-reduced path is $ABBAA$. From the perspective of A and B, *Continuity* forms paths only involving themselves, which we refer to as "AB-reduced" paths. By only counting consecutive encounters, we hypothesize that *Continuity* $= 1$ may increase the specificity of the detection of a true referral and hence meaningful relationship in exchange for lower sensitivity (more true collaborations missed). Therefore, if hospitals with higher coordination of care involve more immediate referrals and fewer distinct physicians, imposing *Continuity* conditions may increase the difference in the network structure between highly- and lowly-coordinated hospitals. Because it outputs AB-reduced path(s) only involving physicians A and B, simplifying evaluation of the remaining factors, *Continuity* is a natural first step in processing referral paths.

2.2 Revisit

The *Revisit* factor is motivated by the conjecture that "closed loops" of referral are a stronger marker of a meaningful relationship, or at least of an influence-spreading opportunity, than α with no loops. *Revisit* specifies the fundamental unit (or pattern) for quantifying the strength and thus the existence of an edge. After first applying *Continuity*, the two fundamental units we consider for quantifying the strength of the AB-edge in each α are AB and ABA. When *Revisit* $= 0$, a visit to A followed by a visit to B in α leads to a contribution to the edge from A to B and AB is the fundamental unit. When *Revisit* $= 1$, a return-visit to A following the visit to B must occur for α to contribute and ABA is the fundamental unit. For example, if *Continuity* $= 0$ the "AB-reduced" path of $\alpha = ACBCAA$ is $ABAA$ and contributes to the AB-edge irrespective of *Revisit*. However, because the ABA sequence occurs but the BAB sequence does not, if *Revisit* $= 1$ then α only contributes to the AB-edge. Because *Multiplicity* depends on *Revisit*, the latter cannot be processed as a standalone step. We discuss *Revisit* further in Sections 2.3.1 and 2.3.3.

2.3 Multiplicity

Multiplicity encodes the strength of a relationship between A and B in α. By virtue of its reliance on the fundamental unit for accumulating edge information, *Multiplicity* interacts directly with *Revisit*. As noted in Section 2.2, the fundamental unit when *Revisit* $= 0$ is the subsequence AB and when *Revisit* $= 1$ the fundamental unit is the subsequence ABA. The strength of the relationship is quantified according to five distinct measures (levels) and recorded as the value $\mathcal{X}_{AB}$ in the physician network adjacency matrix:

- *Existence* (f) – indicator variable recording the existence of a fundamental relationship of A and B in α.
- *Total-Count* (g) – record of the total number of unique fundamental relationship units from A to B in α ignoring repeated visits to the same physician; $g = 0$ if the fundamental relationship unit does not occur in α.
- *Total-Ordering* (h) – record of the total number of fundamental relationship units from A to B in α summing all occurrences of the fundamental unit that abide to the direction of the edge; $h = 0$ if the fundamental relationship unit does not occur in α.

Clearly, *Total-count* and *Total-ordering* increase in expectation with the length of α, making them reflective of care volume. Therefore, to obtain counterpart measures that are not as systematically related to the volume of care, it is also useful to have "normalized" versions of the measures, the form of which we discuss in Section 2.3.2. We break those out separately:

- *Normalized Total-Count* (g^*) – normalized record of the total number of fundamental relationship units of A and B in α.
- *Normalized Total-Ordering* (h^*) – normalized record of the total number of fundamental relationship units from A to B in α summing all occurrences of the fundamental unit that abide to the direction of the edge.

We will make use of the following notation: for referral path α and physician A occurring in α, let $r^A(\alpha)$ denote the list of visit indices to physician A in α. The first visit receives a 1. For example,

$$\text{if } \alpha = ABBACB \text{ then } \begin{cases} r^A(\alpha) &= (1,4) \\ r^B(\alpha) &= (2,3,6) \\ r^C(\alpha) &= (5) \end{cases} \tag{1}$$

We let $n_A(\alpha)$ denote the number of occurrences of physician A on referral path α. In the above example, $n_A(\alpha) = 2$.

Now for the computation of *Multiplicity*.

2.3.1 Case 1: Revisit = 0

1. *Existence*: If referral path α contains a visit to A before at least one visit to B set $f_{AB}(\alpha) = 1$, otherwise $f_{AB}(\alpha) = 0$.
2. *Total-count*: This records the total number of times on referral path α that a visit to A is followed by a visit to B, ignoring repeated visits with the same physician. Denote the number of such occurrences by $g_{AB}(\alpha)$. For example, the sequence $\alpha = ABBBAB$ yields $g_{AB}(\alpha) = 2$ (and $g_{BA}(\alpha) = 1$).
3. *Total-ordering*: For each visit to A in α, count the number of visits to B that follow and sum over that count. More precisely, let

$$h_{AB}(\alpha) = \sum_{i,j} I(r_i^A < r_j^B) \tag{2}$$

where the indices for i and j denote the visit number (allowing consecutive visits to the same physician) and I denotes the standard indicator function; $I(\mathrm{e}) = 1$ if the expression e is true and 0 otherwise. For example, $\alpha = ABBBAB$ yields $h_{AB}(\alpha) = 5$ (and $h_{BA}(\alpha) = 3$). The quantity in (2) is related to the Mann-Whitney u-statistic (Mann and Whitney, 1947).

Total-count ignores runs of visits to the same physician while *total-ordering* accounts for runs of multiple visit to the same physician by comparing the chronological order of each visit to physician A with that of each visit to physician B. Visits on the same day may be accounted by assuming that the true order of visits is the order in which the claims are listed.

2.3.2 Normalized Levels under Revisit = 0

The normalized versions of *total-count* and *total-ordering* are obtained by dividing by $w^g_{AB}(\alpha) = n_A(\alpha) + n_B(\alpha)$ and $w^h_{AB}(\alpha) = n_A(\alpha)n_B(\alpha)$, respectively, to obtain:

$$\begin{aligned} g^*_{AB}(\alpha) &= w^g_{AB}(\alpha)^{-1} g_{AB}(\alpha) \\ h^*_{AB}(\alpha) &= w^h_{AB}(\alpha)^{-1} h_{AB}(\alpha). \end{aligned}$$

The normalized quantities reflect relative patient flow from physician A to B on referral path α relative to the maximum possible given $n_A(\alpha)$ and $n_B(\alpha)$. Because $w^g_{AB}(\alpha) = w^g_{BA}(\alpha)$ and $w^h_{AB}(\alpha) = w^h_{BA}(\alpha)$ it follows that $\max(g^*_{AB}(\alpha) + g^*_{BA}(\alpha)) = 1$ and $\max(h^*_{AB}(\alpha) + h^*_{BA}(\alpha)) = 1$. That is, the maximum sum of the normalized *total-count* and *total-ordering* for α is a sum of proportions that sum to 1.

Whereas $g_{AB}(\alpha)$ and $h_{AB}(\alpha)$ will generally increase with the length of a referral path and allow longer referral paths to have more influence on the network, $g^*_{AB}(\alpha)$ and $h^*_{AB}(\alpha)$ lie within the (0, 1] interval and consider each α to be of equal information content regardless of its length.

2.3.3 Case 2: Revisit = 1

Accounting for a "revisit" encodes the assumption that a physician is only meaningfully influenced by another physician if a patient referred to another physician returns to the initial referrer for a follow-up appointment. This construct is encoded in making the sequence ABA (rather than AB) the fundamental unit of physician relationships. (Note that even if the fundamental unit is ABA, the associated record is in the (A, B) entry of the adjacency matrix, reflecting a directed referrer-referee relationship.) The levels of *Multiplicity* are modified accordingly:

1. *Existence*: Each referral path is given a binary indicator of whether or not ABA occurred. That is, set $f_{ABA}(\alpha) = 1$ if true, 0 otherwise.
2. *Total-count*: The number of times the subsequence ABA occurs on α, collapsing over repeated visits to the same physician. For example, if $\alpha = ABBBABA$ then $g_{ABA}(\alpha) = 2$ (and $g_{BAB}(\alpha) = 1$).
3. *Total-ordering*: A rank-based measure that accounts for *Multiplicity* in the visits to A and B when quantifying the frequency of ABA subsequences. Unlike *Total-Count*, repeat visits are enumerated. Specifically:

$$h_{ABA}(\alpha) = \sum_{\ell,i,j} I(r_\ell^A < r_i^B < r_j^A), \tag{3}$$

the product of the number of repeated visits first to A, then B, and finally to A again. Equivalently, for each AB subsequence in α, count of the number of visits to B that follow and sum over that count. For example, the values of *total-ordering* for $\alpha = AAABBBA$ and $\alpha = AAABBBAAA$ are:

$$h_{ABA}(AAABBBA) = 9 \text{ and } h_{ABA}(AAABBBAAA) = 27$$

Normalized versions of $g_{ABA}(\alpha)$ and $h_{ABA}(\alpha)$ are obtained by dividing by their maximum possible values given $\{g_{AB}(\alpha), g_{BA}(\alpha)\}$ and $\{h_{AB}(\alpha), h_{BA}(\alpha)\}$, respectively. The respective divisors are $w_{ABA}^g(\alpha) = g_{AB}(\alpha) + g_{BA}(\alpha)$ and $w_{ABA}^h(\alpha) = h_{AB}(\alpha)h_{BA}(\alpha)$, although $w_{ABA}^h(\alpha)$ is only applied to referral paths with $h_{AB}(\alpha)h_{BA}(\alpha) > 0$ as it otherwise equals 0. We denote these additional levels by $g_{ABA}^*(\alpha) = w_{ABA}^g(\alpha)^{-1}g_{ABA}(\alpha)$ and $h_{ABA}^*(\alpha) = w_{ABA}^h(\alpha)^{-1}h_{ABA}(\alpha)$. Analogous to Section 2.3.2, $\max(g_{ABA}^*(\alpha) + g_{BAB}^*(\alpha)) = 1$ and $\max(h_{ABA}^*(\alpha) + h_{BAB}^*(\alpha)) = 1$. Because $w_{ABA}^f(\alpha) = 1$ the addition of the normalized measures only introduces two additional levels of *Multiplicity*, bringing the total to 5.

2.4 Summing over Referral Paths

Finally, the entry $\mathcal{X}_{AB}$ must account for the contributions from all referral paths including patient visits to both A and B. If $k_{AB} \in \{f_{AB}, g_{AB}, h_{AB}, g_{AB}^*, h_{AB}^*\}$ denotes a *Multiplicity* level, then the entry $\mathcal{X}_{AB}$ will be the sum of k_{AB} over all α:

$$\mathcal{X}_{AB} = \sum_{\alpha} k_{AB}(\alpha) \tag{4}$$

2.5 Post-processing: Directed and Binary

After forming a weighted-directed network by applying *Continuity*, *Revisit* and *Multiplicity*, we may leave the network as is, convert it to an undirected form, convert it to binary, or convert it to undirected and binary. These transformations augment the 20 weighted-directed network settings of the projection to a physician network with 60 additional settings.

Although there are multiple ways of obtaining an undirected network from a directed network, here we simply sum the weights in each direction. That is, we compute $\mathcal{X}_{AB}^{\text{un}} = \mathcal{X}_{AB} + \mathcal{X}_{AB}^T$. The undirected networks serve as controls for the corresponding directed networks when evaluating the value of retaining directional information.

The network is converted to binary by applying a threshold rule to $\mathcal{X}_{AB}$ or $\mathcal{X}_{AB}^{\text{un}}$ (and assigning a 1 or 0 to a given entry depending on whether or not it meets the threshold). In our motivating example, the threshold was the 20'th percentile of the edge-strength distribution for the corresponding weighted network. Edges with larger values become equal to 1 while all other edges are 0 (i.e., are null edges). If the weight at the 20'th percentile is 0 (i.e., fewer than 20% of edges are non-null), all positively-valued edges in the weighted network form the edges in the corresponding binary network.

3 Evaluating the Performance of the Projection Methods

The potential to learn from hospital networks about factors associated with the adoption of health procedures and protocols depends on (i) the extent to which the networks vary and (ii) the association of the network features with the adoption outcome(s). We define the "discriminatory power" of a (one-mode) projection with respect to a quantitative network feature in terms of the heterogeneity of that feature across the networks. The focus on heterogeneity is supported by the fact that in simple linear regression the greater the variance of a predictor the greater the information available to detect its relationship with outcomes. The "predictive power" of a projection with respect to an outcome is the extent to which it informs the prediction of that outcome. Discriminatory power and predictive power measure (i) and (ii), respectively.

3.1 Discriminatory Power

In the following, we assume that the network features are comparable to each other in terms of scale. For example, in the motivating application, we normalize the measures to have a marginal variance of 1 across the hospitals and projections (Section 4). Let Z_{kij} be a random variable denoting the normalized network feature k evaluated on hospital i using projection j and X_j a vector whose elements indicate levels of the factors for projection j. Observed values of Z_{kij} are denoted z_{kij} and the sample mean and variance of $\{z_{kij}\}_{i=1:n}$ by $\bar{z}_{kj}$ and s_{kj}^2, respectively. The discriminatory power specific to feature k for projection j is s_{kj}^2 and the optimal projection for feature k is $\text{opt}_k = \max_{j=1:J}\{s_{kj}^2\}$.

We use three network-wide measures of discriminatory power. Two are the average sample variance across the features, $\bar{s}_j^2 = K^{-1}\sum_{k=1}^{K} s_{kj}^2$, and the associated average rank of the sample variance across the features, $\bar{r}_j = K^{-1}\sum_{k=1}^{K} r_{kj}$, where r_{kj} is the rank from largest to smallest of s_{kj}^2 (rank J is assigned to the projection with the largest s_{kj}^2) in the set $\{s_{1j}^2, \dots, s_{Kj}^2\}$. The rationale for $\bar{r}_j$ is that it is more outlier resistant than $\bar{s}_j^2$. For the third, we evaluated the proportion of variation that occurs between the hospitals when features are treated as repeated measurements and analyzed using the statistical model

$$Z_{kij} = \beta_{0j} + \theta_{kj} + \epsilon_{kij} \tag{5}$$

where $\theta_{kj} \sim \text{Normal}(0, \tau_j U_j)$ and $\text{cov}(\epsilon_{1ij}, \dots, \epsilon_{Kij}) = \sigma_j^2 R_j$ for each j, where U_j and R_j are correlation matrices. The standardization of the network measures across the observed data prior to model estimation makes the assumption underlying (5) of homogeneous hospital and residual variances across the measures justified. One measure of the proportion of the total variation in the network measures that occurs between the networks is $\rho_j = \tau_j^2/(\tau_j^2 + \sigma_j^2)$, which corresponds to a ratio of traces of the between-hospital to the sum of the between- and within-hospital covariance matrices. Unfortunately, the elements of U_j cannot be identified. To overcome this challenge, we fit the special case of the model in which $U_j = R_j = I$, the identity matrix, equating ρ_j with the intraclass correlation coefficient (ICC). Given the approximate nature of this calculation, we view the resulting estimates of ρ_j, denoted $\hat{\rho}_j$, as a model-based counterpart to $\bar{s}_j^2$ and $\bar{r}_j$.

To quantify the importance of the projection factors to discriminating between networks we regress s_{kj}^2 on X_j for each network feature $k = 1, \ldots, K$:

$$s_{kj}^2 = X_j^T \beta_k + \epsilon_{kj}. \tag{6}$$

As long as the sample variances are evaluated over a large enough sample (e.g., up to 2,219 hospitals in the motivating example), the central limit theorem ensures that the empirical distribution of s_{kj}^2 and of ϵ_{kj} in (6) will be close to normal and thus that a linear model and normal-based inference is appropriate. Furthermore, in the motivating example, we restrict to networks whose largest connected component contains > 20 physicians, helping to make the network features well-defined (Section 4). In general, an alternative to normal-based inference would be to use a bootstrap to generate estimates of standard errors and evaluate statistical inferences.

To evaluate which factors explain the most overall variation between the projections, we regress $\hat{\rho}_j$, $\bar{sz}_j = \bar{s}_j^2/\bar{z}_j$ and $\bar{r}_j$ on X_j using analogous linear models to (6). The reason for the use of $\bar{sz}_j$ as a dependent variable is that a linear variance-mean relationship was evident between s_{kj}^2 and $\bar{z}_{kj}$, a consequence of the restricted scales of the features that are proportions or correlations. Therefore, $\bar{sz}_j$ is a more comparable measure.

3.2 Predictive Power

We define the predictive power of network features Z_{ij} with respect to a network-level outcome Y_i in terms of the area under the ROC curve of the estimated logistic regression of Y_i on Z_{ij}. In our motivating application, $Y_i = \text{Adopt}_i$ denotes the ith hospital's adoption of the capability to perform implantable cardiac defibrillator (ICD) procedures. Let $\pi_{ij} = \Pr(Y_i = 1 \mid Z_{ij})$, where Z_{ij} is the vector of network features for network i under projection j. Because the predictors vary across the $J = 80$ projections, the conditional probability π_{ij} and its estimator $\hat{\pi}_{ij}$ varies across j despite the dependent variable Y_i and the marginal probability $\Pr(Y_i = 1)$ being invariant to j. The general form of the systematic part of the statistical model is

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_j^T Z_{ij} \tag{7}$$

We fit (7) to obtain the vector of estimates $\hat{\beta}_j$ of the coefficients of the network features under projection j and summarize the predictive accuracy of the fitted model by the c-statistic $c_j = \Pr(\hat{\pi}_{hj} > \hat{\pi}_{lj} \mid h \in \{h : \text{Adopt}_h = 1\}, l \in \{l : \text{Adopt}_l = 0\})$, the probability that a randomly selected hospital with ICD capability has a greater estimated probability of adoption than a randomly selected hospital without ICD capability, a non-parametric estimator of the area under the receiver operating characteristic curve (AUC). The inclusion of network size in the vector of network features Z_{ij} in (7) accounts for the possibility that larger hospitals may have more within-hospital referrals and greater likelihood of adopting technologies, allowing the independent association of other features with ICD capability to be estimated.

We are particularly interested in the extent to which network features that extract directional information from referral paths improve the predictive accuracy of the model over that attained without using directional information. We define the weighted and binary baseline projections to be the projections to an undirected network using the existence level of *Multiplicity* without imposing *Continuity* or *Revisit*. Let Z_{ij}^B denote the vector of network features for the baseline network counterpart of the projection that generated Z_{ij}. The added value of the information about directionality in Z_{ij} is quantified by comparing the fit of the logistic regression model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_j^T Z_{ij}^B + \lambda_j^T Z_{ij} \tag{8}$$

to that of the model that only includes the elements of Z_{ij}^B as predictors. The value of the directional information in patient referral paths may be represented by the absolute and the percentage difference of the c-statistic, denoted $\Delta(c)$ and $\%\Delta(c)$ respectively, and the difference in the deviance statistics between the two models.

4 Motivating Example, Network Measures, Data, and Data Wrangling

Implantable Cardiac Defribrillators (ICD) can provide major clinical benefit by alleviating symptoms and preventing heart attacks (Moss et al., 2002; Kadish et al., 2004; Bardy et al., 2005; Al-Khatib et al., 2011). However, they are expensive and may activate and shock a patient when not needed. Hence, overuse should be avoided and the specification of who is an appropriate patient to receive an ICD is debated. The association of network measures with the ICD adoption status of hospitals may provide important insights into why some hospitals adopt and why others do not adopt ICD capability (Moen et al., 2016, 2018, 2019).

The physician shared-patient networks for each hospital are developed using Medicare claims data from 2010 involving 2,219 hospitals in the USA. In accordance with the most restrictive threshold used in the publicly available Centers for Medicare and Medicaid Services (CMS) data on the number of patients referred between pairs of physicians (CMS, 2021), we set $t_{\max} = 30$ days as the maximum time to allow for consecutive visits to be included in a referral path α.

The hospital-specific networks are defined by assigning each physician to the hospital where their office was located, if they worked in a hospital, or otherwise the hospital where the plurality of their patients went for overnight hospital stays (Bynum et al., 2007). Therefore, each physician appears in a single network. Because network features may be ill-defined or highly variable on small networks, we excluded hospitals with 20 or fewer nodes in the largest connected component of the network under any setting of the factors to obtain the sample of 2,219 hospitals (approximately one-third the total number of hospitals in the USA).

To enable the diffusion of ICDs to be investigated, we used the patient ICD registry of procedures performed in 2011 to determine the number of ICD procedures performed by hospital with a hospital defined as an adopter if they performed at least one ICD procedure (the use of registry data from the subsequent years ensures that no data is used to both define the network and measure ICD status). Although the availability of the ICD registry and their importance in medicine makes the diffusion of ICDs a compelling example, there are many other procedures and technologies whose diffusion might also be studied using this approach (e.g., coronary-artery stents and other surgical interventions for abdominal aortic aneurysm or carotid-artery disease).

We use a broad range of features to evaluate the 80 sets of hospital networks formed by the 80 one-mode projections. The features are grouped in Table 1 by whether they are base features – those defined for undirected networks and thus for all networks – or are specific to

Table 1: Network-level summary features.

Measure	Definition
Degree	The total edge-weight of a node to or from other nodes
Base features (defined on undirected networks)	
Density	The average edge-weight across the network
Centralization	The variance of the degree distribution of the network
Triangles (undir)*	Proportion of (undirected) triangles that are closed
Clustering Coef*	Proportion of triangles involving a node that are closed averaged across nodes
Size*	Number of nodes or physicians in the largest connected component (LCC) of the network
Isolates*	Proportion of nodes with 0 non-null edges
Components*	Number of distinct components in network
Diameter*	Shortest path length (number of edges) across largest connected component
Directed Features (defined only on directed networks)	
ExpansePop	Correlation of in- and out-degree
Assortativity	Within-dyad correlation of degree: Four variants in directed networks
Reciprocity	Within-dyad correlation of edge weights
Transitive triads	Proportion of triads that are transitive ($A \to B$, $B \to C$, $A \to C$) among non-null triads
Cyclical triads	Proportion of triads that are 3-cycles ($A \to B \to C \to A$) among non-null triads

**Invariant to weighted edges by virtue of treating them as binary*

directed networks. The remaining base features are invariant to whether the network is weighted or binary. (See Wasserman and Faust (1994) and Newman (2010) for a thorough description of network statistics.) The directed features are undefined for undirected networks and, therefore, endow directed networks with additional information compared to undirected networks.

For greater variance to imply greater discriminatory power, network features must have the same scale across the projection methods. We applied self-standardizing transformations to density and centralization and then normalized all features to have a marginal variance of 1 across the hospitals and projections in the study sample (Section 1 of Supplemental Appendix).

We explored whether weighting the observations by the number of nodes in the network further justified the assumption of homogeneous variances in (6) but found that it didn't appreciably and so used unweighted regression models to estimate the contributions of the projection factors to discriminatory power. The models in (7) and (8) for evaluating predictive power were also unweighted.

5 Results

We report results on the relationship of the projection methods and their underlying factors to the discriminatory power of the network with respect to each network feature and the network as a whole in Section 5.1. This is followed by results that compare the predictive accuracy of a hospital's network structure with their ICD status across the projection methods in Section 5.2.

5.1 Results: Discriminatory Power of Projections for Hospital Networks

The values of opt_k, X_{opt_k}, and $s^2_{k\text{opt}_k}$ for the $k = 1, \ldots, 16$ features evaluated on the physician shared-patient networks within 2,219 US hospitals are first presented in Table 2. The z-statistics for the estimates of β_k of the feature-specific and of the pooled features ρ_j (the ICC), $\bar{sz}_j$, and $\bar{r}_j$ on the projection factors for the model in (6) are discussed in the text with details presented in Section 2 of the Supplemental Appendix. The top 10 directed projections in terms of ρ_j are then presented in Table 3.

5.1.1 Optimal Feature-specific Projections

The greatest discriminatory power for density, centralization, the correlation between in- and out-degree, the four assortativity features, and reciprocity occurred when both *Continuity* and *Revisit* were binding and weighted edges were retained (see first two base feature and first 6 directed feature rows of Table 2). The optimal projections for these features show a clear advantage over their second best projection; the optimal projection for density is the most definitively identified (37.6% superiority over second-best projection) with reciprocity and centralization the second- and third-most. The optimal settings of *Multiplicity* for these features were always other than existence; e.g., inter-hospital variation in density and centralization was greatest with total-ordering while variation in reciprocity was greatest with normalized total-ordering. These findings imply that the frequencies of the sequences A and B, AB and BA, and ABA and BAB across all α contain valuable information beyond simply knowing that physicians A and B were visited at least once by the same patient.

Because the remaining base features are undirected and binary in nature (transitivity and clustering coefficient are functions of undirected open and closed triangles, size and proportion of isolates are based on counts of actors, number of components and diameter are enumerated assuming edges exist or not), the *Multiplicity* and *Binary* factors are not relevant to them. However, *Continuity* and *Revisit* have a substantial impact with their presence associated with more heterogeneity in most features between the networks. The most notable exception is the size of the largest connected component, for which relaxing *Continuity* and *Revisit* was optimal.

The results for size can be explained by the fact that the more lenient the requirement for an edge to exist, the larger the expected size of the largest connected component and, by virtue of size being a count variable, the greater its variability. In contrast, the proportion of isolates is greatest under the most restrictive conditions for forming the network (*Continuity* and *Revisit* imposed); raising the bar for an edge to exist thereby increased the chance that a given physician has no edges. Because the expected number of components and diameter increase as the density of the network decreases, it makes sense that applying constraints that reduce density increases their expectation and variability.

Unlike other directed network measures, the proportions of transitive and cyclical triads attain their greatest heterogeneity with both *Continuity* and *Revisit* non-binding (Table 2, bottom of second segment). An explanation of the former is that not allowing $C \neq A, B$ to occur

Table 2: Feature-specific Optimal Projection and Discriminatory Power.

	Design factors					Improve
Measure	*Mult*	*Cont*	*Rev*	*Bin*	Het	2nd (%)
			Base Features			
Density	TOrder	1	1	0	7.032	37.6
Centralization	TOrder	1	1	0	11.469	13.6
Triangle (undir)*	.	1	0	0	0.405	0
Clustering Coef*	.	1	1	0	0.403	0
Size*	.	0	0	0	1.673	0
Isolates*	.	1	1	0	0.161	0
Ncomponents *	.	1	1	0	2.736	0
Diameter*	.	1	1	0	1.387	0
			Directed Features (only defined on directed networks)			
Cor(In,Out −Deg)	TOrder	1	1	0	1.690	7.30
Assort(In,In)	TCount	1	1	0	1.760	7.84
Assort(In,Out)	TCount	1	1	0	1.831	3.72
Assort(Out,In)	TOrder	1	1	0	1.831	11.8
Assort(Out,Out)	TCount	1	1	0	1.698	8.06
Reciprocity	NCount	1	1	0	0.937	18.8
Transitive Triads	NCount	0	0	1	1.224	0.06
Cyclical Triads	NCount	0	0	1	1.283	0.05
			Network-wide (Pooled feature) optimization over directed networks			
ICC (ρ_j)	TCount	1	1	0	0.093	22.9
Ave Stand Var ($\bar{sz}_j$)	TOrder	1	1	0	2.80	46.9
Ave Rank ($\bar{r}_j$)	TOrder	1	1	0	29.7	3.93

**Feature invariant to weighted edges. Heterogeneity (Het) corresponds to $s^2_{\text{kopt}_k}$ for the 16 features and to the value of the quantity in the left-hand column for the three network-wide measures. Improve 2nd (%) denotes superiority of the optimal projection over the second-best projection in terms of a percentage. The five levels of Multiplicity (Mult) are abbreviated by: Existence = Exist, Total-count = TCount, Total-ordering = TOrder, Normalized total-count = NCount, and Normalized total-ordering = NOrder. The remaining three factors are abbreviated as Continuity = Cont, Revisit = Rev, and Binary = Bin.*

between visits with A and B removes an important source of transitive triads. In another departure from the prevailing results for directed features, triadic features had more heterogeneity if the network was transformed to binary as opposed to remaining as weighted.

The network-wide measures are in agreement on all factors other than *Multiplicity* with total-count yielding higher $\hat{\rho}_j$ while total-ordering yielded higher average values of $\bar{sz}_j$ and $\bar{r}_j$.

5.1.2 Projection Factor Effects and Top Ten Projections

The results for the projection factor regression coefficients in model in (6) reveal that *Continuity* and *Revisit* are generally highly significant (z-statistics almost always > 5 and often > 10) and that the direction of their estimated coefficients (positive when enforced, negative when not enforced), while the estimated effects of *Multiplicity* and *Binary* are generally less pronounced

Table 3: Optimal Overall (Across Feature) Directed Projections in terms of ICC.

Projection factors				Hospital variances		Inter-hospital ICC	
Mult	*Cont*	*Rev*	*Bin*	Between	Within	All	Directed Only
TCount	1	1	0	0.102	0.989	0.093	0.233
Exist	1	1	0	0.074	0.897	0.076	0.199
TCount	1	1	1	0.061	0.784	0.073	0.167
TOrder	1	1	1	0.061	0.784	0.073	0.167
NCount	1	1	1	0.061	0.784	0.073	0.167
NOrder	1	1	1	0.061	0.784	0.073	0.167
Exist	1	1	1	0.061	0.784	0.072	0.167
TOrder	1	1	0	0.143	1.925	0.069	0.187
Exist	0	1	1	0.030	0.456	0.061	0.169
NCount	0	1	1	0.030	0.456	0.061	0.169

Results are for the model given by the variant of Equation (5) in which ρ_j (the ICC for the model in Equation 5) is the common residual correlation between all pairs of features. Note the simplifying abbreviations for the five levels of Multiplicity = Mult (Exist, TCount, TOrder, NCount, and NOrder) and Continuity = Cont, Revisit = Rev, and Binary = Bin.

(Section 2 of Supplemental Appendix). These findings are consistent with the optimal projections in Table 2.

We estimated interaction effects by augmenting (6) with second and higher-order terms involving the projection factors. The predominant interaction effect was between *Continuity* and *Revisit*, with significant positive effects on ρ_j and $\bar{sz}_j$ (z-statistics of 4.02 and 4.59, respectively – results not reported) and on the standardized scaled-variance of several network features (density, specialization, and most assortativity measures). While there were sporadic significant interactions between other factors, including one third-order interaction, these were insufficiently prevalent to warrant reporting. The fact that the presence of both *Continuity* and *Revisit* often had a greater effect than either factor alone is an important finding and motivates examining the factor settings of the best performing projection methods (Table 3).

The top ten directed projections in terms of ICC are dominated by the enforcement of *Continuity* and *Revisit* (Table 3). All 10 impose *Revisit* and the top 8 impose *Continuity* $= 1$. Consistent with the ICC results in the bottom segment of Table 2, the optimal projection to a one-mode directed physician network used the total-count level of *Multiplicity* (Table 3, first row). This projection yielded $\hat{\rho}_j = 0.093$ when evaluated over all features and $\hat{\rho}_j = 0.233$ when evaluated over only the directed features. Both of these optimal values are well above the estimate of ρ_j for the second-best projection. The fact that all five binary-network projections were tied as the third-equal optimal projection with $\hat{\rho}_j = 0.073$ reflects that, in general, *Multiplicity* has minimal impact on the discriminatory power of binary networks.

5.2 Results for Predictive Power

Table 4 presents the value of directional information and related quantities for the top five weighted and top five binary projections with respect to $\Delta(c)$. Among weighted network projections, imposing *Revisit* with the normalized total-ordering level of *Multiplicity* but without re-

Table 4: Optimal projections for maximizing the value of directional network information.

Projection factors			AUC (c-statistic)				Deviance
Mult	*Cont*	*Rev*	Base	All	$\Delta(c)$	$\%\Delta(c)$	difference
			Weighted				
NOrder	0	1	0.581	0.602	0.021	25.6	29.5
NOrder	1	1	0.581	0.599	0.018	22.4	33.3
NOrder	1	0	0.581	0.599	0.017	21.3	26.4
NCount	1	0	0.581	0.598	0.017	20.7	22.2
NCount	0	1	0.581	0.598	0.017	20.3	26.1
			Binary				
TCount	1	0	0.590	0.607	0.016	18.3	34.1
TOrder	1	0	0.590	0.606	0.016	17.5	33.4
NCount	1	0	0.590	0.605	0.015	16.7	32.2
NOrder	1	0	0.590	0.605	0.015	16.2	31.6
Exist	1	0	0.590	0.605	0.014	16.0	32.4

AUC = area under the ROC curve (AUC), also known as the c-statistic. The Base and All c-statistics are from the models in Equations (7) and (8), respectively. $\%\Delta(c)$ and the Deviance difference are computed using the fits of these two models; bigger values indicate greater improvement in model fit. Note the simplifying abbreviations for the five levels of Multiplicity = Mult (Exist, TCount, TOrder, NCount, and NOrder) and Continuity = Cont, Revisit = Rev, and Binary = Bin.

quiring *Continuity* obtained the greatest $\Delta(c)$ of 0.021 – a percentage increase of $\%\Delta(c) = 25.6\%$ – over the base (undirected) projection. The improvement is well above that for the second-best projection ($\Delta(c) = 0.018$, $\%\Delta(c) = 22.4\%$), which had the greatest reduction of deviance (33.3), and differed from the top projection in that *Continuity* was imposed. A feature of the top-five ranked projections is that *Multiplicity* is always one of its normalized levels, suggesting that standardizing the contribution from each referral path to the edge weight enhanced predictive accuracy. The sixth best weighted projection (not shown) is the directed version of the base projection for weighted networks (existence level of *Multiplicity*, no *Continuity* or *Revisit* constraints). It's $\%\Delta(c)$ of 19.4% may be thought of as the value of capturing the simplest form of directionality and the difference of 6.2% between this value and the value of 25.6% for the optimal projection as a measure of the added-benefit of more nuanced processing of directional information.

The top five $\%\Delta(c)$ projections to a binary one-mode physician network enforce *Continuity* and relax *Revisit* (lower segment of Table 4). These projections include the five levels of *Multiplicity* with total-count the best level. The same projection is optimal in terms of maximum reduction of deviance. The consistency of these results with respect to *Continuity* and *Revisit* consolidates the finding that their enforcement and relaxation, respectively, characterize the best projection to a one-mode binary network. *Multiplicity* appears less relevant to an analysis involving binary as opposed to weighted networks.

For a comparison of the estimated regression coefficients between the Optimal and Baseline Estimated Models of ICD-status, see Section 3 of the Supplemental Appendix.

6 Discussion

In this paper we developed and explored novel unipartite projections of the standard bipartite network that connects physicians and their patients. This work was motivated by the historical disregard shown to the time-order of patient visits with their physicians in the construction of shared-patient networks and makes use of the underlying referral path information (An et al., 2018b,a, 2019) that is lost (or neglected) in the formation of the traditional patient-physician bipartite network. We found that referral paths contain substantial information that can be used to better distinguish the one-mode networks from one another thereby enhancing the potential to discover relationships of network features to important health variables compared to what is possible with networks based on currently-used undirected projections. Imposing the *Continuity* and *Revisit* constraints led to greater heterogeneity in all network features related to directionality. The total-count and total-ordering levels of *Multiplicity* led to networks that were more heterogeneous than under the commonly-used existence rule. Although not always the case, the retention of weighted edges often led to greater heterogeneity in important network features between the resulting networks.

In our study of the hospital-level adoption of implantable cardiac defribrillators (ICDs) we showed that extracting directional information and placing more weight on referral paths with certain features (e.g., feedback loops and direct referrals) increased the predictive accuracy of shared-patient networks. Imposing *Revisit* (i.e., making ABA the core subsequence or building-block for the AB edge in the network) and the normalized total-ordering level of *Multiplicity* increased the c-statistic by 25.6% over the predictive accuracy of the standard undirected weighted network. The fact that the predictive accuracy of hospital-level ICD-adoption status was improved by the addition of directed network information confirms that time-ordered referral paths contain valuable information above and beyond undirected bipartite networks even without knowing the true relationship statuses. While motivated by the development of directed physician networks from health insurance claims data, the methodology in this paper may be applied to any situation in which a directed network is to be constructed from the sequence of encounters of one type of actor with actors of another type.

There are several ways in which the current work can be extended. Consideration of the sequence ABA as a stronger marker of a meaningful relationship is a starting point for a more general study that addresses questions such as: Do longer sequences such as $ABABA$ hold even greater significance than quantified by total-count and total-ordering (or their normalized variants)? Should the physicians be further labeled by specialty or subspecialty? Rather than summing referral paths over the entire study period to define edges and thus a single network for each study unit (e.g., hospital), an alternative would be to allow edges to change status with time. The involvement of dynamic networks would make research on temporal paths relevant (Armbruster et al., 2017). Another extension is to use physician speciality labels in conjunction with information on directionality to obtain networks characterized by particular types of referrals between types of specialists or to enlarge the projection space over which to further improve the discriminatory and predictive power of the resulting networks. One may also question the utility of projecting the network to a unipartite space at all. Network methods exist that analyze data in the bipartite space, avoiding the loss of information that occurs in projection, such as hypergraph representations (Gerow et al., 2015). However, for many uses of networks, such methods will be further afield than using well-known methods on a network constructed a different way.

By applying similar methods to those described in this paper, projections for optimizing actor (e.g., physician) positional network features could also be estimated. Each actor (e.g., physician in a hospital network) may be characterized by a vector of actor positional network features that could be used for tasks such as predicting a physician's future practice of medicine (e.g., whether they'll adopt certain medical therapies). Each physician could also be characterized by the commonality of their role in referral paths. Physicians or hospitals who generally initiate the same sequence of visits might practice medicine in a very rigid or organized way, which could be good for minimizing variations in health care but bad for innovation. Hamming distances or other measures of diversity could be used to quantify the similarity of the sequences initiated at each physician. In addition, the referral paths involving groups of physicians may be quantified by the extent to which they reflect cohesion in their coordination of care delivered to patients.

An important factor underlying referral paths is how long an interval to use to determine when one referral path ends and another begins. While we based our choice of $t_{\max} = 30$ on the value used by CMS in their physician shared-patient datasets CMS (2021), $t_{\max}$ could be optimized over. Such optimization could be accomplished by using cross-validation to find the most predictive $t_{\max}$ simultaneously with the optimal combination of projection factors using one part of the data with the rest of the data used to evaluate performance (discriminatory or predictive power); this is another avenue for future work.

We hope this paper increases interest in using claims data to extract directional information about physician relationships. Such information may enhance comparative analyses involving physician shared-patient networks by generating more informative networks and summary network features for use in subsequent analyses. More broadly, we hope that the recognition of the diversity of explanatory power encoded in the choices of factor (projection-parameter) settings as well as the utility of bringing to bear the framework of factorial experiment or other designed experiments proves useful across many different disciplines and their associated network studies.

Supplementary Material

In the supplemental appendix, which accompanies this manuscript and will be published on the journal website in the "Supplementary Material" section, we present expanded descriptions of the data wrangling and additional results that were not able to be included in the main paper due to length restrictions. The results in the supplemental appendix are supported by the text that would have accompanied them in the main text had space permitted. In addition, we also refer readers to the following GitHub site to obtain R and Python code used in the analyses: https://github.com/kiwijomalley/OptimalBipartiteProjection.

Acknowledgement

The authors thank Peter Mucha and Feng Fu of the Department of Mathematics at Dartmouth College and the P01 research group at The Dartmouth Institute for helpful feedback.

Funding

This work was supported by the National Institutes of Health in the USA (grant number P01 AG019783).

References

Al-Khatib SM, Hellkamp A, Curtis J, Mark D, Peterson E, Sanders GD, et al. (2011). Non-evidence-based icd implantations in the United States. *Journal of the American Medical Association*, 305: 43–49.

An C, O'Malley AJ, Rockmore DN (2018a). Referral paths in the U.S. physician network. *Applied Network Science*, 3(1): 20.

An C, O'Malley AJ, Rockmore DN (2019). Towards intelligent complex networks: The space and prediction of information walks. *Applied Network Science*, 4(1): 35.

An C, O'Malley AJ, Rockmore DN, Stock CD (2018b). Analysis of the U.S. patient referral network. *Statistics in Medicine*, 37(5): 847–866.

Armbruster B, Wang L, Morris M (2017). Forward reachable sets: Analytically derived properties of connected components for dynamic networks. *Network Science*, 5(3): 328–354.

Bardy GH, Lee KL, Mark DB, Poole JE, Packer DL, Boineau R, et al. (2005). Amiodarone or an implantable cardioverter-defibrillator for congestive heart failure. *New England Journal of Medicine*, 352: 225–237.

Barnett ML, Landon BE, O'Malley AJ, Keating NL, Christakis NA (2011). Mapping physician networks with self-reported and administrative data. *Heath Services Research*, 46: 1592–1609.

Bynum JP, Bernal-Delgado E, Gottlieb D, Fisher E (2007). Assigning ambulatory patients and their physicians to hospitals: a method for obtaining population-based provider performance measurements. *Health Services Research*, 42(1p1): 45–62.

CMS (2021). Physician shared patient datasets.

Coleman J, Menzel H, Katz E (1959). Social processes in physicians' adoption of a new drug. *Journal of Chronic Diseases*, 9(1): 1–19.

Farine DR, Whitehead H (2015). Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology*, 84.

Foti NJ, Hughes JM, Rockmore DN (2011). Nonparametric sparsification of complex multiscale networks. *PLoS One*, 8(2): e16431.

Gerow A, Lou B, Duede E, Evans J (2015). proposing ties in a dense hypergraph of academics. In: *Social Informatics. SocInfo 2015* (TY Liu, C Scollon, W Zhu, eds.), volume 9471 of *Lecture Notes in Computer Science*. Springer, Cham.

Griliches Z (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica: Journal of the Econometric Society*, 501–522.

Iyengar R, Van den Bulte C, Valente TW (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2): 195–212.

Kadish A, Dyer A, Daubert JP, Quigg R, Estes M, Anderson KP, et al. (2004). Prophylactic defibrillator implantation in patients with nonischemic dilated cardiomyopathy. *New England Journal of Medicine*, 350: 2151–2158.

Liang X, Zhang S, Liu Y, Ma Y (2020). Information propagation formalized representation of micro-blog network based on Petri nets. *Scientific Reports*, 10(657).

Mann HB, Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1): 50–60.

Moen EL, Austin AM, Bynum JPW, Skinner JS, O'Malley AJ (2016). An analysis of patient-sharing physician networks and implantable cardioverter defibrillator therapy. *Health Services and Outcomes Research Methodology*, 16: 132–153.

Moen EL, Bynum JPW, Austin AM, Chakraborti G, Skinner JS, O'Malley AJ (2018). Assessing variation in implantable cardioverter defibrillator therapy guideline adherence with physician and hospital patient-sharing networks. *Medical Care*, 56(4): 350–357.

Moen EL, Bynum JPW, Skinner JS, O'Malley AJ (2019). Physician network position and patient outcomes following implantable cardioverter defibrillator therapy. *Health Services Research*, 54(4): 880–889.

Montgomery DC (2017). *Design and Analysis of Experiments*. John Wiley & Sons.

Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, et al. (2002). Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New England Journal of Medicine*, 346: 877–883.

Newman MEJ (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physicial Review E*, 64(1): 016131.

Newman MEJ (2010). *Networks: An Introduction*. Oxford University Press, Oxford, UK.

Onnela JP, O'Malley AJ, Keating NL, Landon BE (2018). Comparison of physician networks constructed from thresholded ties versus shared clinical episodes. *Applied Network Science*, 3(1): 28.

Pavlopoulos GA, Kontou PI, Pavlopoulou A, Bouyioukos C, Markou E, Bagos PG (2018). Bipartite graphs in systems biology and medicine: A survey of methods and applications [published correction appears in gigascience (2020), 9(1), giz130]. *Gigascience*, 7(4): giy014.

Rogers EM (2010). *Diffusion of Innovations*. Simon and Schuster.

Ryan B, Cross N (1943). Acceptance and diffusion of hybrid corn seed in two Iowa communities. *Rural Sociology*, 8: 15–24.

Schwarz AJ, McGonigle J (2011). Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. *NeuroImage*, 55(3): 1132–1146.

Skinner JS, Staiger DO (2005). Technology adoption from hybrid corn to beta blockers. *National Bureau of Economic Research, Working Paper*, 11251: http://www.nber.org/papers/w11251.

Valente T (1995). *Network Models of the Diffusion of Innovations*. Hampton Press, New Jersey.

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