Bayesian Network Motifs for Reasoning over Heterogeneous Unlinked Datasets

Yi Sui[1*], Alex Kwan[1], Alexander W. Olson[1], Scott Sanner[1] and Daniel A. Silver[2]

[1*]Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, M5S 3G8, ON, Canada.
[2]Sociology, University of Toronto, 5 King's College Road, Toronto, M5S 3G8, ON, Canada.

*Corresponding author(s). E-mail(s): suiyiamy@gmail.com;
Contributing authors: alex.kwan@mail.utoronto.ca;
alex.olson@mail.utoronto.ca; ssanner@mie.utoronto.ca;
dan.silver@utoronto.ca;

Abstract

Modern data-oriented applications often require integrating data from multiple heterogeneous sources. When these datasets share attributes, but are otherwise unlinked, there is no way to join them and reason at the individual level explicitly. However, as we show in this work, this does not prevent probabilistic reasoning over these heterogeneous datasets even when the data and shared attributes exhibit significant mismatches that are common in real-world data. Different datasets have different sample biases, disagree on category definitions and spatial representations, collect data at different temporal intervals, and mix aggregate-level with individual data. In this work, we demonstrate how a set of Bayesian network motifs allows all of these mismatches to be resolved in a composable framework that permits joint probabilistic reasoning over all datasets without manipulating, modifying, or imputing the original data, thus avoiding potentially harmful assumptions. We provide an open source Python tool that encapsulates our methodology and demonstrate this tool on a number of real-world use cases.

Keywords: Data Science, Bayesian Networks, Unlinked Data

1 Introduction

A broad range of research fields have gained new interest in combining traditional data sources with readily available, inexpensive, large-scale data sources. This desire to work with big data can introduce both opportunities and problems not previously seen by researchers in those fields. Datasets frequently suffer from data sparsity, sampling bias, poor granularity, or simply not having all attributes of interest. For example, surveys increasingly suffer from low response rates, social media fails to account for demographics which have been slow to adopt these platforms, national surveys are unable to provide finer level city estimates, and tax records lack non-reported income.

Therefore, it is sometimes necessary to integrate multiple datasets when a single dataset alone is insufficient. With the rise of the open-source data movement, there are many publicly available data sources that can be used to supplement a dataset to resolve such issues. Unfortunately, it can be difficult to jointly reason over heterogeneous data merged from multiple sources in a robust way.

For data merging, we describe data as broadly falling into one of three types: linked, unlinked, and aggregate-level. Linked data means that a foreign key exists to identify a record across multiple datasets uniquely. In contrast, unlinked data may have common attributes among multiple datasets, but no foreign key exists, often due to privacy concerns or the promise of anonymity. Finally, while linked and unlinked data often record individual samples, aggregate-level data is calculated or categorized over populations [1]. This type of data is readily available when individual record-level data is hard to obtain due to privacy concerns (e.g. census) or when tracking individuals is difficult (e.g. insect counts).

While linked data can be merged using existing SQL-like tools, there is not yet a universally accepted technique to merge unlinked or aggregate-level datasets. Since novel Big Data sources frequently fall into these latter categories, it is increasingly necessary to create a reliable mechanism for merging these types of data sources that vary in quality and suitability.

In this work, we present a discrete probability framework that allows one to perform probabilistic reasoning over multiple heterogeneous datasets in the absence of foreign keys. This framework utilizes Bayesian Network motifs (i.e., patterns) to handle common data mismatches including statistical (sampling bias) mismatch, individual vs. aggregate-level record mismatch, and attribute domain mismatch. Statistical mismatch occurs when datasets are drawn from different sampling distributions. Individual vs. aggregate-level mismatch refers to merging individual records with aggregate-level data. Finally, attribute-domain mismatch occurs when the same attribute is represented differently across multiple datasets.

The framework is incorporated into a Python library, *ppandas*, which extends the popular data science package *pandas*. In presenting our approach both as a methodology and as a flexible, easy-to-use scientific package, we hope to allow for both the expansion of this framework by other researchers, as

well as for quick deployment of these interpretable techniques within existing Python data analysis pipelines.

We begin in Section 2 by highlighting the limitations of existing approaches to merge unlinked and aggregate-level datasets. Section 3 then outlines the real-world datasets and motivating use cases that contain the common data mismatches explained in Section 4. These issues are then addressed by the framework methodology in Section 5. In Section 6, we apply the framework in a series of experiments based on the motivating use cases. Finally, Section 7 discusses the framework's applicability and potential future work.

2 Literature Review

Previous work has proposed that inference across multiple tables in databases can be performed by assuming there is a global schema mapping for all tables [2–5]. Each table contains a subset of attributes from the global schema, and inference can be performed by querying over the global schema. However, it is assumed that each object in the real world is associated with a unique identifier, such as a social insurance number of a person. Therefore, these methods can not work with unlinked data.

Probabilistic Relational Models (PRMs) [6, 7] were proposed as a formalism for leveraging Bayesian network inference for probabilistic reasoning over databases. A PRM specifies a template for a probability distribution over a database. The template includes a relational component that describes the relational schema for the interested domain and a probabilistic component that describes the probabilistic dependencies that hold in the interested domain [8]. However, PRMs do not explicitly focus on merging datasets and only work with linked data.

Probabilistic Databases [9, 10] approach merging unlinked datasets by representing all attribute pairs in the database as tuples associated with a single attribute — the probability of its presence in the dataset. While Probabilistic Databases can be captured as graphical models, there are some distinctions. Firstly, the probabilistic model in Probabilistic Databases consists of a set of independent tuples, whereas graphical models allow one to represent causal relationships and dependencies between attributes compactly. Moreover, Probabilistic Databases do not explicitly address mismatches between datasets and instead focus on integrating data uncertainties within existing relational database frameworks.

Record linkage, also known as entity resolution, is explicitly intended to recover links in unlinked data. These methods attempt to link and merge records across datasets representing the same entity in the absence of foreign keys. One such method is PRL (probabilistic record linkage), [11] which calculates a similarity score [12, 13] between pairs of records to assess prospective matches. For example, a combination of name, address, birth date can be sufficient to identify individuals in two survey datasets [11]. Another record linkage method is Bayesian linkage. Bayesian linkage calculates the posterior

probability that two records refer to the same entity [14]. However, record linkage methods are unsuitable when one does not have sufficient information to discover unique linkages or when there are no entities common to both two datasets. Furthermore, record linkage can raise privacy concerns, and just a small amount of false matches could potentially lead to severely biased outcomes [15]. In our work, we employ Bayesian Network Motifs to allow probabilistic inference across datasets. Therefore, we do not attempt to recover linkage between datasets and thus can answer queries across sampling distributions from disjoint populations.

Statistical matching [16], also known as data fusion, is a method of combining datasets through common attributes such as demographic information to obtain a fused dataset. However, such methods do not compactly represent a joint distribution using a graphical model structure and thus can suffer from data sparsity. A popular way of performing statistical matching is to impute the non-overlapping attributes as missing data on a per-record basis. The imputation is performed through machine learning methods such as regression by learning the relationship between overlapping attributes (e.g. demographic data) and missing attributes in the dataset [17–19]. However, imputation does not provide a probabilistic distribution over all possible outcomes and instead provides a deterministic classification. Additionally, this method does not solve the issue of how to combine two or more complete data sources that have identical attribute sets.

Another domain of related work is *multiple frame methods* [20], in which there are several datasets with the same schema that each contain a different sampling distribution for a subset of the entire population. By taking the union of all the sampling distributions, one can estimate the distribution for the entire population. For example, the same survey could be given to cell phone and landline users to obtain a better understanding of the entire population of phone users while accounting for the fact that some users have both a landline and cell phone [21]. In multiple frame methods, statistics in all datasets are concatenated to a final frame with duplicates removed. The population total of an attribute can be estimated as the weighted average from each frame [20, 22] in order to adjust for duplicates. However, ideal multiple frame methods require upfront sampling methodologies and can only merge datasets that have identical attributes that were collected with the same procedures. These methods are also not suitable for datasets where the frame membership is unclear.

A final vein of related work concerns the *learning of probabilistic models from heterogeneous (possibly inconsistent) local sources of data* in the framework of gradient optimization for tractable models [23] or Bayesian methods for learning Bayesian Network parameters [24] and structure [25] from multiple local data sources. One very recent work in this vein also attempts to recover causal structure and local interventions in the underlying data sources [26]. While all of these works provide significant advances for the learning of TPMs and Bayesian Networks from local data sources and share common motivations

with our work, these works diverge from our setting in two key defining aspects: (1) these existing works assume that the underlying datasets have the same attribute (variable) definitions, whereas a core motivation of our work is to explicitly resolve mismatches in attribute values (different numerical intervals, different spatial regions, different categorical values) and aggregate-individual mismatch; (2) these existing works attempt to blend the distributions from the underlying sources whereas our approach aims to leverage one distribution as the reference distribution (identified by the end user) and to "correct" the other distributions to align with the reference distribution.

To achieve our goal of reasoning over unlinked data from multiple local data sources, we build on our own prior work [27] that introduced a Bayesian network framework for this task, but which did not address Bayesian Network motifs to handle various types of mismatch that is the core focus of this article. *To the best of our knowledge, no previous work proposes methods to address mismatch issues between heterogeneous datasets explicitly, including statistical (sampling bias) mismatch, individual vs. aggregate-level record mismatch, and attribute-domain mismatch that we contribute in this work.*

3 Data Description and Motivating Use Cases

Before we proceed to specific technical issues, we first describe real datasets that support the motivating use cases for our work. Each dataset has two attribute types: independent and dependent. Independent attributes are not affected by any other attributes, while dependent attributes are influenced by the independent attributes. For example, in a survey, the demographic attributes are independent attributes, and question responses are dependent attributes. In the following schema tables, attributes in bold font indicate independent attributes. We include our pre-processing pipeline for each dataset within the *ppandas*[1] repository.

3.1 Datasets

Toronto Election Study (TES)

The TES [28] is a survey that polled 3,000 Toronto residents from Sept 27, 2014 to Oct 27, 2014, before the 2014 Toronto mayoral election. The survey collected demographic attributes, policy opinions, and voter preferences. The schema of the TES data after data pre-processing is shown in Table 1.

CBC Vote Compass Toronto 2014 (VC)

CBC Vote Compass [29] is a public online interactive tool that asked 33,352 participants to provide their opinion on a variety of policy issues and their candidate preference for the 2014 Toronto mayoral election. After completing the questionnaire, participants were told which candidate best aligned with

[1] `https://github.com/D3Mlab/ppandas/experiments`

Table 1 TES data schema

Attribute	Value
Age	{(17,23], (23,28], ... (58,63], (63,114]}
City ward	{Ward 1,Ward 2... Ward 44}
Employment status	{Working, Family caring, Student, Retired, Unemployed, Other}
Birthplace	{Canada, Not Canada, Unknown}
Increase immigration	{For, Against}
Bike Lanes	{For, Neutral, Against}
Candidate preference (pre-election)	{John Tory, Doug Ford, Olivia Chow, Other}

their policy beliefs, which could have impacted each participant's actual vote. The data is proprietary and can not be released publicly. The schema of the VC data after pre-processing is shown in Table 2.

Table 2 VC data schema

Attribute	Value
Age	{(17,23], (23,28], ... (58,63], (63,114]}
City ward	{Ward 1,Ward 2... Ward 44}
Employment status	{Working, Family caring, Student, Retired, Unemployed, Other}
Birthplace	{Canada, Not Canada, Unknown}
Helping existing immigrants to adapt	{For, Against}
Bike Lanes	{For, Neutral, Against}
Candidate preference	{John Tory, Doug Ford, Olivia Chow, Other}

Since the Vote Compass was designed as a utility for voters first, and as a survey second, it did not follow a specific sampling methodology. As a result, it is not representative of the city's population, with considerable skew towards younger Toronto residents who live in the central downtown area.

2011 Census Toronto Ward Profiles

Toronto's entire 2011 population of 2,615,090 is broken down in the Canadian census [30] by Age and Ward as shown in Table 3. Note that the 2011 Census does not provide information regarding employment status or birthplace.

Table 3 Census data Schema

Attribute	Value
Age	{[17,19], [20,24], [25,29], [30,34] , [35,39], [40,44], [45,49], [50,54], [55,59], [60,64], [65,114]}
City ward	{Ward 1,Ward 2... Ward 44}
Counts	Count of population

Instead, this information was collected in the voluntary 2011 National Household survey [30], which is regarded as less reliable than the mandatory Census and, therefore, we chose to exclude these two demographic attributes.

Chicago West Nile Surveillance 2007-2013

A Kaggle competition dataset [31] provides Chicago's surveillance program for the West Nile Virus for the years 2007, 2009, 2011 and 2013. Every week from late-May to early-October, public workers set up traps located throughout the city and tested mosquitoes for the virus. For these four years, there are 10,507 trap test records. Each trap test result includes data, as shown in Table 4.

Table 4 West Nile Virus dataset schema

Attribute	Value
Date	Date of the record
Latitude	Latitude of the trap's geo-location.
Longitude	Longtitude of the trap's geo-location.
Number of mosquitoes	{1,...,50}
West Nile virus present	{Yes, No}

Note that not all the traps are tested every week, and records only exist when mosquitoes are found at a certain trap on a certain date.

Chicago Community Data Snapshot March 2014

Chicago contains 77 community areas that are used by the city government as statistical units for a variety of socioeconomic indicators [32]. This dataset's attributes of interest are shown in Table 5.

Table 5 Chicago Community Data Schema

Attribute	Value
Community area	{Rogers Park, West Ridge,..., Edgewater}
Median income (USD)	{13,345, 110,365}

Chicago Community Area Boundaries

This shapefile [33] contains the geographical boundaries for Chicago's 77 community areas.

3.2 Motivating Use Cases

We now present four scenarios that exhibit various types of data mismatch which motivate our work. They are presented now in order to familiarise the reader with the context in which we intend for our tool to be valuable. Due to the inherent uncertainty of such predictions (i.e., in unlinked data, we can

only make predictions about typical entities with the observed properties), we seek to obtain a probability for each behavioural outcome.

(1) 2014 Toronto Mayoral Election

In order to provide more accurate voting forecasts, a social scientist seeks to correct the TES distributions for age group and city ward with their respective Toronto census distributions. While the TES sampling adheres to an age quota, the Toronto census is an aggregate-level dataset that provides a much more accurate representation of the voting population's age and city ward distribution. Also, since the other TES demographic attributes (birthplace and employment status) do not have a respective Toronto census distribution, they are excluded from this use case. The social scientist proceeds to integrate the two data sources but realizes the census' age groups differ from those used in TES.

(2) Toronto Bike Lanes

While TES and VC come from two separate sources and have different demographic distributions, a policymaker seeks to make an informed decision on the installation of bike lanes by reasoning jointly over both surveys. More specifically, they seek to understand how demographic attributes influence one's stance on bike lanes. As given previously, both datasets provide a surveyor's age group, city ward, employment status, birthplace, and their bike lanes stance. Except for age, all attributes and the bike lane stance question exhibit no mismatch, meaning for each question, both surveys' responses can be mapped to a common set.

The policymaker believes that TES's distribution of demographic attributes is a more representative sample since VC is viewed as a convenience sample where large amounts of responses were collected cheaply and swiftly through an open, online platform. However, TES's small sample size means it does not have a sample for every permutation of age group, city ward, employment status, and birthplace. Therefore, using VC can provide many more samples to more accurately learn the joint distribution over all the attributes. They do not believe naively stacking the two datasets is the right approach since VC's larger sample size would dominate TES. Most crucially, to respect the surveyors' privacy, neither survey allows the policymaker to uniquely identify individuals meaning no foreign key exists to join the datasets with a unified global schema.

(3) Chicago West Nile

An epidemiologist seeks to combine the Chicago West Nile dataset with Chicago's community area statistics. By doing so, the epidemiologist can explore the relationships between West Nile Virus and socioeconomic measures such as median income. The epidemiologist seeks to use GIS coordinates of mosquito traps to create a Voronoi diagram [34] such that all points in a

Voronoi cell are closer to the mosquito trap associated with that cell than any other mosquito trap. In this instance, the plane is the unified boundary of Chicago's 77 community areas and the n points are the trap locations. The epidemiologist seeks to allocate portions of each community area to the Voronoi diagram polygons based on regions of overlap.

(4) Toronto Attitude Towards Immigration

As given in subsection 3.1, the TES records a participant's stance towards increasing immigration to Canada, and VC records one's stance on whether Toronto should help existing immigrants to adapt to their new life. A social scientist is interested in investigating whether these two beliefs are related. The scientist thus poses the following research question: if one believes that Toronto should help new immigrants to adapt, are they more or less likely to be against increasing immigration to Canada?

To answer this question, the social scientist would need to combine responses from TES and VC. However, there is no unique identifier for each respondent in TES or VC. Additionally, the populations that the two sets of respondents were drawn from are different.

4 Issues when merging heterogeneous datasets

As mentioned previously, joining linked datasets with foreign keys is straightforward using SQL-like tools. For example, Table 6 illustrates two surveys performed on the same participants, which can be viewed as homogeneous datasets. Each participant is associated with an id, with Survey 1 providing a participant's age and Survey 2 providing their stance towards installing additional bike lanes. Since the "id" field is a unique identifier, the two tables are linked. Thus by performing a join, one can match a participant's id to both their age and bike lane attitude to obtain the joint distribution table. After obtaining the table, one can answer a wide array of queries. For example, one can calculate the probability that a participant is against installing more bike lanes given they are older than 40 by using the joint table: $P(Bike\ Lane = Against | Age > 40) = \frac{1}{2}$.

Table 6 Survey 1 (left) and Survey 2 (center) are linked datasets that can be joined using the foreign key "id" to obtain the joint distribution (right) of both age and bike lane stance.

id	Age
01	20
02	25
03	63
04	45

id	Bike Lane
01	For
02	For
03	For
04	Against

id	Age	Bike Lane
01	20	For
02	25	For
03	63	For
04	45	Against

However, in this section, we discuss the many issues that can arise when attempting to join heterogeneous unlinked and aggregate-level datasets.

4.1 Unlinked data with shared independent attributes

When joining multiple unlinked datasets, there is no unique identifier to act as a foreign key, but one can leverage shared attributes. For example, in the "Toronto Attitude Towards Immigration" use case provided in subsection 3.2, both the TES and VC lack an "id" attribute but share the independent attribute of a participant's age. Additionally, the TES asked about participants' stances towards increasing immigration while VC has information on stances towards helping immigrants adapt. The sample data is shown in Table 7.

Table 7 Two unlinked datasets with shared independent attributes: sample data of TES (left) asked about age and stance towards increasing the amount of immigration in Canada; sample data of VC(right) asked about age and if they agree with helping new immigrants adapt in Toronto. VC participants were younger than TES participants.

Age	Increase immigration
(23,28]	Against
(28,33]	Against
(38,43]	Against
(43,48]	Against
...	...

Age	Help immigrants to adapt
(23,28]	For
(28,33]	For
(23,28]	Against
(28,33]	Against
...	...

However, when joining the two datasets to discover the relationship between one's attitude towards increasing immigration and helping new immigrants to adapt, we cannot naïvely map the records by the Age attribute to obtain the joint table since it is not a unique identifier. Therefore, unlike with linked datasets, it is impossible to match all of the records in the two datasets. Another issue is that TES and VC inquired about different individual participants. They represent different sampling distributions, exhibiting statistical (sampling-bias) mismatch.

4.2 Unlinked data with shared independent and dependent attributes

If we have two datasets from different sampling distributions with shared independent and dependent attributes, it is often of interest to utilize both datasets to obtain a comprehensive understanding of the dependent attributes. For example, the "Toronto Bike Lanes" use case given in subsection 3.2 stated that TES and VC both contain attributes of age and stances towards bike lanes. With these two datasets, the policymaker intends to have a better understanding of people's stances towards bike lanes.

One method is to stack the two datasets together naïvely. However, it raises the concern that the combined dataset could be biased towards the younger, downtown population. In this scenario, the policymaker is interested in the overall Toronto population, which is similar to the TES population. While one could just use the single TES dataset, one may which to leverage the rich VC data since it contains ten times the number of samples.

Table 8 Two unlinked datasets from different sampling distributions with shared independent and dependent attributes: sample data of TES (left) and VC (right). They both have information on a participants' age and their attitude toward building more bike lanes.

Age	Bike Lanes
(23,28]	For
(28,33]	For
(38,43]	For
(43,48]	Against
...	...

Age	Bike Lanes
(23,28]	For
(28,33]	For
(23,28]	For
(28,33]	For
...	...

4.3 Individual vs. aggregate-level record mismatch

In the era of big data, datasets are commonly represented with aggregated counts as opposed to individual entries. When trying to join aggregate-level datasets with other data sources, we face issues not only with the absence of a foreign key but with individual vs. aggregate-level record mismatch as well.

For example, in the "2014 Toronto Mayoral Election" use case given in subsection 3.2, the census data contains aggregate-level records, whereas TES data contains individual records. Sample data is shown in Table 9. To illustrate the concept of individual vs. aggregate-level record mismatch, we temporarily ignore the fact that the 2011 Toronto Census and TES employed different age groups.

Table 9 Individual vs. aggregate-level record mismatch: sample data of 2011 Toronto Census (left), which contains the aggregated counts of the city population's age and location distributions. Sample data of TES (right) contains the same information as well as candidate preference but on an individual record basis.

Age	**Location**	Counts
(23,28]	Ward 11	2000
(28,33]	Ward 1	4000
(33,38]	Ward 28	1000
(38,43]	Ward 28	1000
(43,48]	Ward 40	2000
...	...	...

Age	**Location**	Candidate preference
(23,28]	Ward 1	Olivia Chow
(28,33]	Ward 20	Other
(38,43]	Ward 17	Other
(43,48]	Ward 22	Other
...	...	...

To obtain a better estimation for the outcome of the mayoral election, the social scientist seeks to combine the city demographics knowledge from the aggregate-level census data with TES's individual record candidate preferences.

4.4 Attribute-domain mismatch

In addition to the previously stated issues, heterogeneous datasets may also have different representations for the same attribute. We denote this issue 'attribute-domain mismatch', and when trying to join datasets with attribute-domain mismatch, one faces the problem of how to unify these different representations.

Attribute-domain mismatch can be further broken into three types: numeric interval mismatch, arbitrary categories mismatch, and spatial representation mismatch.

4.4.1 Numeric interval mismatch

When the shared attributes between two datasets are represented with different numerical ranges, we denote it as numeric interval mismatch. One frequent attribute that exhibits numeric interval mismatch is Age. For example, in the "2014 Toronto Mayoral Election" use case given in subsection 3.2, the census data represents age in bins: {[17,19], [20,24], [25,29], [30,34] , [35,39],[40,44], [45,49], [50,54], [55,59], [60,64], [65,114]}, while TES represents age in bins: {(17,23], (23,28], (28,33], (33,38],(38,43], (43,48], (48,53), (53,58], (58,63], (63,114]}. Sample data of this mismatch is provided in Table 10.

Therefore, when the social scientist attempts to combine the information from the two datasets, in addition to aggregate-level data handling for the census, there is a need to handle the numeric interval mismatch for the age attribute.

Table 10 Two unlinked datasets with numeric interval mismatch: sample data of census (left) and TES(right). Both datasets contain information on age and location. TES also has information on candidate preference. The age bins employed in the two datasets differ.

Age	**Location**	Counts
[25,29]	Ward 11	2000
[30,34]	Ward 1	4000
[35,39]	Ward 28	1000
[35,39]	Ward 28	1000
[40,44]	Ward 40	2000
...	...	...

Age	**Location**	Candidate preference
(23,28]	Ward 1	Olivia Chow
(28,33]	Ward 20	Other
(38,43]	Ward 17	Other
(43,48]	Ward 22	Other
...	...	...

4.4.2 Arbitrary category mismatch

When the shared attribute between two datasets are represented with different arbitrary categories, we denote it as arbitrary category mismatch.

An example of arbitrary category mismatch is shown in Table 11. As mentioned in subsection 3.2, both TES and VC contain information on age and attitude toward bike lanes. To illustrate the concept of arbitrary category mismatch, we assume that in TES the attitude towards bike lanes is represented with 5 categories: {"Strongly For", "Moderately For", "Neutral", "Moderately Against", "Strongly Against"} whereas in VC it is represented with 2 categories :{"For","Against"}.

Table 11 Two unlinked datasets with arbitrary category mismatch: sample data of TES (left) and VC (right). They both contain information on age and attitude towards bike lanes. The attitude towards bike lanes is represented with different categories.

Age	Bike Lanes
[20,40)	Strongly For
[20,40)	Moderately For
[20,40)	Neutral
[40,60)	Strongly Against
[40,60)	Moderately Against
[60,80)	Strongly For
...	...

Age	Bike Lanes
[20,40)	For
[20,40)	Against
[20,40)	For
[40,60)	Against
[40,60)	Against
[40,60)	Against
...	...

4.4.3 Spatial Representation Mismatch

When the shared attribute between two datasets is represented with different spatial representations, we denote it as 'spatial representation mismatch'. For example, in the "Chicago West Nile" use case given in subsection 3.2, we seek to merge West Nile virus (WNV) test results with median income across community areas. For simplicity, in this example, we simplify Chicago's boundary into a rectangle that has two community areas and three mosquito traps as provided in Table 12. After creating a Voronoi diagram using the three mosquito traps, Figure 1 illustrates that the polygons representing Voronoi cells and community areas differ.

Table 12 Two unlinked datasets with spatial representation mismatch: Chicago West Nile surveillance (left) and Chicago community areas (right) use different Polygons that cannot be straightforwardly mapped to each other, as shown in Figure 1.

Voronoi Cell	WNV Present
Trap A	+
Trap A	-
Trap B	-
Trap B	-
Trap C	+
Trap C	+
...	...

Community Area	Median Income
Community Area 1	$31,000
Community Area 2	$56,000

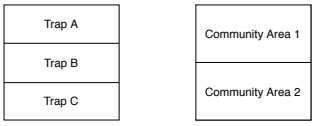


Fig. 1 The spatial representation used in Table 12: Chicago West Nile Surveillance divides the area into 3 Voronoi cells, and Chicago community area boundaries divide the area into two community regions.

When both spatial representations are regions, we refer to this as a region-region mismatch. Geo-locations can also be represented as point coordinates,

for example, latitude and longitude. Therefore, region-point and point-point mismatch may also occur in practice.

5 Methodology

5.1 The Bayesian Network Representation

Before exploring the framework, it is necessary to introduce fundamental Bayesian network concepts and terminology.

Model Definition A Bayesian network [35] is defined as an "acyclic directed graph in which nodes represent random variables and arcs represent directed probabilistic dependencies among them". They provide a compact representation of a joint probability distribution by exploiting conditional independence assumptions. Formally, for a finite set of discrete random variables, $X = \{X_1, ..., X_n\}$ the joint distribution is equal to the factorized product:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1}) = \prod_{i=1}^{n} P(X_i|pa(X_i)) \tag{1}$$

where $pa(X_i)$ denotes the set of parent nodes with edges directed into each child node X_i and each factor $P(X_i|pa(X_i))$ stores a conditional probability distribution (CPD) in tabular form that contains every possible combination of variable assignments with maximum likelihood parameters estimated from the data's empirical distribution. For parentless nodes, $P(X_i|pa(X_i) = \emptyset)$ is known as the prior probability $P(X_i)$. (1) satisfies the Markov condition where the probability of each variable given its parents is independent of its non-descendants in the DAG.

Exact Inference Given a Bayesian Network, we can infer any probability query $P(Y|E = e)$ that consists of: query variables $Y \subset X$ conditioned on evidence variables $E \subseteq X$ with instantiation e, where $Y \cap E = \emptyset$. We employ the Variable Elimination algorithm which efficiently eliminates non-query variables to analytically compute the posterior probability distribution $P(Y|E = e)$. Note that if $E = \emptyset$ then $P(Y)$ is denoted as the marginal distribution over random variables Y.

Independence Relationships For a given Bayesian network with 3 variables where A is the shared variable, there are only 4 possible DAG structures as illustrated in Fig 2.

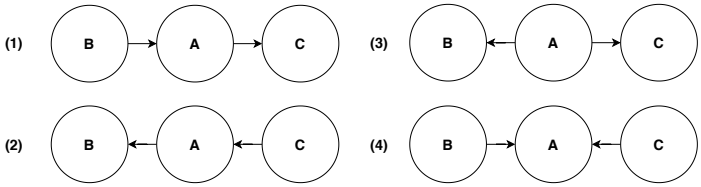


Fig. 2 All possible orientations of 3 nodes where A is the shared variable

The first 3 structures all encode the same I-equivalence (independence-equivalence) relationship: $B \perp C|A$, meaning that B is independent of C if A is given, with no active trail of probabilistic influence existing between C and B. However, if A is not given then $B \not\perp C$, and there is an active trail between C and B. However, the V-structure shown in (4) encodes a different I-equivalence relationship: $B \not\perp C|A$ and when A is not given then $B \perp C$. These properties will be crucial in order to define legal models within the framework.

5.2 Formal Problem Definition and Proposed Framework

In this section we start with basic framework definitions building on our prior Bayesian Network foundations for reasoning over unlinked data [27]. Suppose there exists a *global relation* (tabular data) we wish to model, but rather than a single global relation, we only have multiple unlinked projections of this global relation, which we call *local relations* (also tabular data). For example, if we performed a join over all three tables in Table 6, we would obtain a global relation, but alternately if we dropped the "id" column from all three tables in Table 6, we would obtain three unlinked local relations for this global relation. In addition, local relations may have different joint distributions and can be seen as being drawn according to a randomized subsampling procedure of the global relation. These local relations may also exhibit various types of data mismatch as discussed previously.

More formally, let $R_G(X)$ represent the global relation over a finite attribute set $X = \{X_1, ..., X_n\}$ where each attribute, X_i can be viewed as a random variable. And let $R_{L_j}(X_j)$ for $j = 1, .., k$ represent the j^{th} local relation out of k local relations generated from $R_G(X)$ where $X_j \subset X$. While any given subset of X can exist in multiple local relations, we consider the case where no subset of X can be used as a foreign key to uniquely identify rows across multiple local relations.

Our ultimate goal is to reason jointly over the local relations $R_{L_j}(X_j)$ for $j = 1, .., k$ to answer any query $P(Y|E = e)$ as if we were performing inference over the global relation, $R_G(X)$.

We now introduce the framework for reasoning over multiple datasets using a series of Bayesian Network motifs. Note that the formalisms given use just two datasets but the framework can be extended to an arbitrary number of datasets given no assumptions are violated.

Assume we are given two local relations R_{L_1}, R_{L_2} with sample sizes N_1, N_2 that contain attributes $X_1 \cup X_2 = X$. Let each local relation be represented by a local Bayesian Network, BN 1, BN 2. For each local relation j, let the set of independent attributes $X_j^{Ind} \subseteq X$ be represented by parentless nodes and let the set of dependent attributes $X_j^{Dep} \subseteq X$ be represented as child nodes that have incoming edges from all X_j^{Ind} parent nodes, where $X_j^{Ind} \cup X_j^{Dep} = \emptyset$. Figure 3 demonstrates the Bayesian network structure between independent and dependent attributes for a given local relation R_{L_j}.

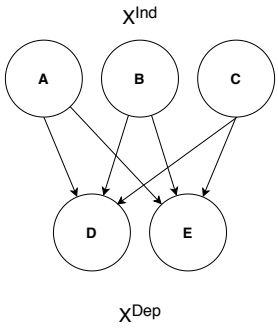


Fig. 3 A Bayesian network with 3 independent variables all having directed arcs leading into both dependent variables.

To obtain numerical parameters for each Bayesian network, maximum likelihood is used which equates to the empirical frequencies given from each local relation. We denote Ref as the reference Bayesian network, representing a single local relation, that contains the probability distributions for all X_j^{Ind} for $j = 1, .., k$. Similarly, we denote Sec as the secondary Bayesian network, representing the other local relation, for which sampling bias is assumed to be restricted only to the independent nodes, and the CPDs between every dependent variable X_j^{Dep} and its parent nodes are assumed to be unbiased with respect to the global relation $R_G(X)$.

As Figure 4 illustrates, if we assume BN 1 is the Ref model and BN 2 the Sec model, we can use the framework to handle data mismatch and create a unified model denoted BN-joined that keeps the probability distributions of X_j^{Ind} from Ref. Note that BN-joined discards the marginal distributions of X_j^{Ind} from Sec and that an independent variable in Sec can be a dependent variable in Ref. Additionally, multiple joins can occur to merge multiple datasets.

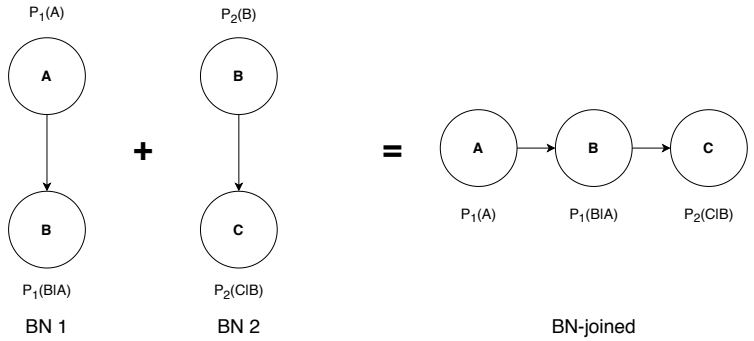


Fig. 4 A Bayesian network is constructed for each of two local relations with variable B being an independent variable in BN 2 but a dependent variable in BN 1. When joining BN 1 and BN2, BN 1 contains the reference distribution, BN 2 contains the secondary distribution, and Variable B is a dependent variable in the resulting BN-joined.

5.3 Framework Assumptions

We now introduce the key assumptions required for using the proposed framework.

As introduced in our prior work [27], parameters in the Bayesian networks can only be learned when all of the independent variables affecting a dependent variable appear in a local relation together. This was discussed in [27] as the definition of "LR-Learnable". In this paper, we implement the data integration method mentioned in [27] in subsubsection 5.4.1. We extend their work by addressing Bayesian networks with shared dependent variables (subsubsection 5.4.2), individual vs. aggregate-level mismatch (subsubsection 5.4.3), and Attribute-domain mismatch (subsubsection 5.4.4).

Our first assumption is that any specified Bayesian network structure must obey the "LR-Learnable" constraint. More specifically, a local Bayesian network can only contain nodes that represent attributes that appear together in a local relation. Furthermore, in order to join two Bayesian networks, they must satisfy the constraint: for each dependent node common to both local Bayesian networks, all of its parent nodes must appear in the reference Bayesian network, Ref.

Additionally, we rely on external expertise to specify the Bayesian network to represent each local relation. This approach is commonly used in practice, especially in medical domains, where experts will often converge on the same causal structure after discussion [36].

Furthermore, we assume that sampling bias is limited solely to the independent variables and that the CPDs between dependent and independent variables are unbiased in all local Bayesian networks. For example, in the "Toronto Bike Lanes" use case stated in subsection 3.2, we assume the sampling bias in VC is because VC focused disproportionately on younger, downtown residents. Thus, Age and City Ward are independent variables that account for VC's biased sampling distribution while the CPD, $P(BikeLane|Age, CityWard, Income, EmploymentStatus, Birthplace)$, is assumed to be unbiased in both the TES and VC Bayesian networks.

Finally, when handling an attribute-domain mismatch for a given variable, if no further information is available, we assume the variable's prior distribution is a uniform distribution to provide an unbiased estimate as subsubsection 5.4.4 will demonstrate. Additionally, only the intersection of attribute domains are handled. For example, in the "2014 Toronto Mayoral Election" use case, when joining TES and Census, the TES only has responses from those aged 17+ due to voting age requirements while the census has data for all ages. Therefore when handling Numeric mismatch for the Age variable, the census data for ages 0 to 16 is discarded. Similarly, for a spatial mismatch between two region sets only the common overlap of the union of the two region sets is handled. Furthermore, for a spatial mismatch between a region set and a point set, only the points from one dataset that is mappable into the other dataset's regions are used. For instance, in the "Chicago West Nile" use case,

the Voronoi diagram's boundary is determined by Chicago's community areas, and any data for mosquito traps that lie outside this boundary are discarded.

5.4 Bayesian Network Motifs

Overall, the procedure for joining two datasets is to first build a Bayesian network for each relation and then join the two Bayesian networks based on shared attributes. After obtaining the joined Bayesian network, probabilistic queries can be performed using inference techniques stated in subsection 5.1.

Table 13 Sample dataset with two attributes: Age and Bike Lanes.

Age	Bike Lanes
20	For
25	For
63	For
45	Against
...	...

For example, Table 13 provides a dataset containing information on age, and stances towards bike lanes. In this dataset, Age is assumed to be the independent attribute and Bike Lanes is the dependent attribute. Therefore, we can build a Bayesian network with parent node Age, child node Bike Lanes, and a directed arc from Age into Bike Lanes. The probability distributions are learned with maximum likelihood, which equals the empirical frequency. The Bayesian network built for Table 13 is shown in Figure 5, where A represents Age and B represents Bike Lanes.

Fig. 5 Bayesian network for Table 13 with the structure determined by external expertise: A represents Age and B represents Bike Lanes

We will now provide a Bayesian network motif to address each issue stated in section 4.

5.4.1 Unlinked data with shared parents

As stated in subsection 4.1, there are two problems joining two unlinked datasets: the absence of a foreign key and statistical mismatch. To address the first issue, we temporarily ignore the fact that TES and VC have different sampling distributions.

Absence of a foreign key: We first construct two Bayesian networks for the two datasets in Table 7, which we denote as BN 1 and BN 2. As shown

in Figure 6, A represents Age, B represents Increase immigration, and C represents Help immigrants to adapt. Since we temporarily assumed the two datasets have the same sampling distribution, $P_1(A) = P_2(A)$.

As illustrated in Figure 6, when joining two datasets with shared parent nodes from the same population, the joined Bayesian network will include the parent node(A) and all children (B and C) from the two local Bayesian networks.

The parameters are kept consistent with the original two Bayesian networks. Therefore, $P(A) = P_1(A) = P_2(A)$. $P(B|A)$ and $P(C|A)$ are obtained from BN 1 and BN 2, respectively.

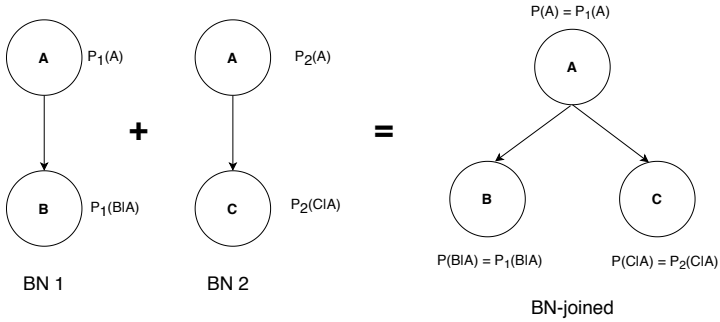


Fig. 6 Unlinked data with shared parents from the same population: the joined Bayesian network, BN-joined contains the shared parent node and both children nodes. The prior distribution of the parent node A can be selected from either BN 1 or BN 2. The conditional probability distributions for the child nodes are selected from their respective local Bayesian networks.

Statistical (sampling-bias) mismatch: Similar to the previous case, we construct two Bayesian networks for the TES and VC data in Table 7, as shown in Figure 7. However, we longer assume $P_1(A) = P_2(A)$ and the Age node in BN 2 is shaded to represent sampling bias as the participants in VC were disproportionately younger. Since TES is believed to contain the *unbiased* (or intended) distribution for Age, BN 1 is assumed to be the Ref model and BN 2 the Sec model.

When performing the join, as shown in Figure 7, the structure of the joined Bayesian network maintains all parent and child nodes from BN 1 and BN 2. For parameters, the parent node distribution comes from the Ref model, BN 1. While the CPDs for each child node are still taken from the original Bayesian networks, $P(B|A)$ and $P(C|A)$ are obtained from BN 1 and BN 2, respectively. In summary, the prior probability parameters should be taken from the Ref Bayes net as justified and discussed in [27].

V-structure limitation: One should be aware that a join cannot always be performed between two Bayesian networks. As mentioned in subsection 5.3, two Bayesian networks BN 1 and BN 2 can only be joined when they satisfy

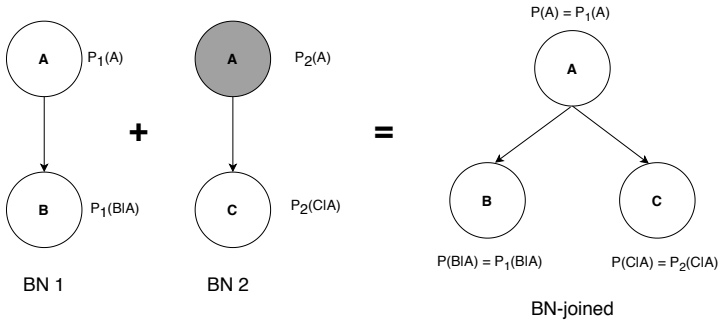


Fig. 7 Unlinked data with shared parents from different populations: the joined Bayesian network, BN-joined, contains the shared parent node and both child nodes. The prior distribution of the parent node A is selected from the *Ref* model, BN 1. The conditional probability distributions for the child nodes are selected from their respective local Bayesian networks.

the LR-Learnable constraint: For every shared child node, all parent nodes must appear in *Ref*.

Figure 8 displays an example of two Bayesian network that cannot be joined. If BN 1 and BN 2 in Figure 7 is built based on expert expertise, where A is dependent on B and C, then the join cannot be performed due to the "V-structure" on node A. In this example, the shared child node A has a parent node C that does not exist in the reference Bayesian network BN 1. Therefore, BN 1 and BN 2 can not be joined, since the conditional probability $P(A|B, C)$ cannot be learned.

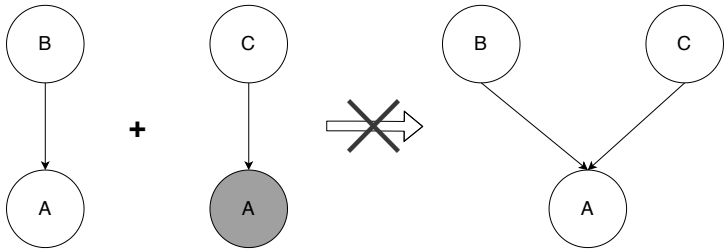


Fig. 8 An example of two Bayesian networks that cannot be joined: V-structure involving node A violates the LR-Learnable constraint.

5.4.2 Unlinked data with shared parents and children from different populations

In this section, we provide the Bayesian network motif that addresses the issue of combining Bayesian networks with statistical mismatch containing shared dependent and independent attributes as mentioned in subsection 4.2; we remark that this case of statistical mismatch was not covered in [27]. First,

we construct two Bayesian networks, BN 1 and BN 2, for the TES and VC data provided in Table 8. As shown in Figure 9, A represents Age and B represents Bike Lanes, where the Age node in BN 2 is shaded to represent VC's sampling bias for that variable. Since TES is presumed to be the unbiased sample, BN 1 is the *Ref* model and BN 2 the *Sec* model.

When joining two Bayesian networks that share both parent and child nodes, the structure of the joined Bayesian network is consistent with the original two local Bayesian networks. Similar to subsubsection 5.4.1 to remove sampling bias mismatch, the prior distribution for the parent nodes is taken from the *Ref* model. To compute CPDs of each shared child node, we can now apply importance sampling to recalibrate probabilities of *Sec* w.r.t. *Ref*; this is simply calculated as the weighted average of the CPD in both the *Ref* and *Sec* models, according to the number of samples in each dataset that the Bayesian networks represent. This weighted average is shown in Figure 9, and the weight denoted as w follows the formula:

$$w = \frac{N_1}{N_1 + N_2}$$

The formal proof of the weighted average formula derived through an importance sampling correction can be found in subsection A.1.

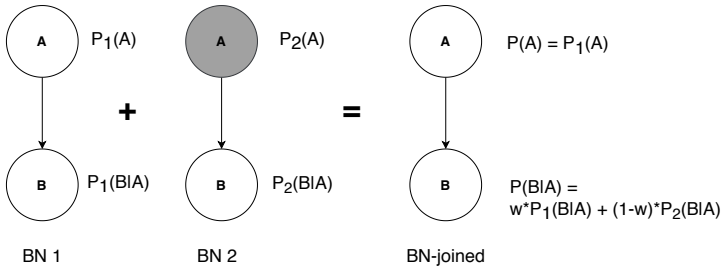


Fig. 9 Unlinked data with shared parents and children from different populations: the structure of the joined Bayesian network, BN-joined, is consistent with BN 1 and BN 2. The prior distribution of the parent node A is selected from *Ref* model, BN 1. The CPD of child node B is calculated as the weighted average of the CPD from BN 1 and BN 2.

5.4.3 Individual vs. aggregate-level Record Mismatch

As stated in subsection 4.3, aggregate-level type sources include the census, where each record does not represent an individual but rather a summary statistic for a group of individuals.

The process of building a Bayesian network for these datasets is similar to if one had individual records. The structure is still based on expert expertise, but the parameters are instead learnt directly from the aggregated counts or percentages.

For example, to handle the census data shown in Table 9, we construct a Bayesian network, BN 1 as shown in Figure 10. BN 1 contains two nodes where A represents Age and B represents Location. Since Age and Location are both independent attributes, BN 1 contains only parent nodes with no conditional dependencies. For the TES dataset, BN 2 is constructed which contains parent nodes for the two previously specified independent attributes in addition to a child node C which represents the dependent attribute, Candidate Preference. To learn parameters for these two models, BN 1 uses the census's already provided aggregated counts while BN 2 uses the empirical counts of the individual records. Lastly, the join can be performed by following the procedure in subsubsection 5.4.1 with BN 1 as the *Ref* model.

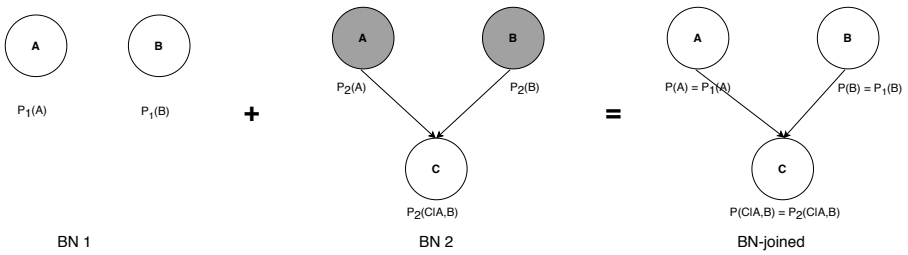


Fig. 10 Individual vs. aggregate-level record mismatch with shared parents: the joined Bayesian network, BN-joined, selects the distributions of the parent nodes from the *Ref* model, BN 1.

To demonstrate building a Bayesian network for aggregate-level dataset with a parent and child node, let us assume that TES records are represented as aggregated counts as shown in Figure 11. The Bayesian network learned for Table 14, therefore, has the same structure as Figure 9 with the parameters for $P(A)$ and $P(B|A)$ learnt through the aggregated counts.

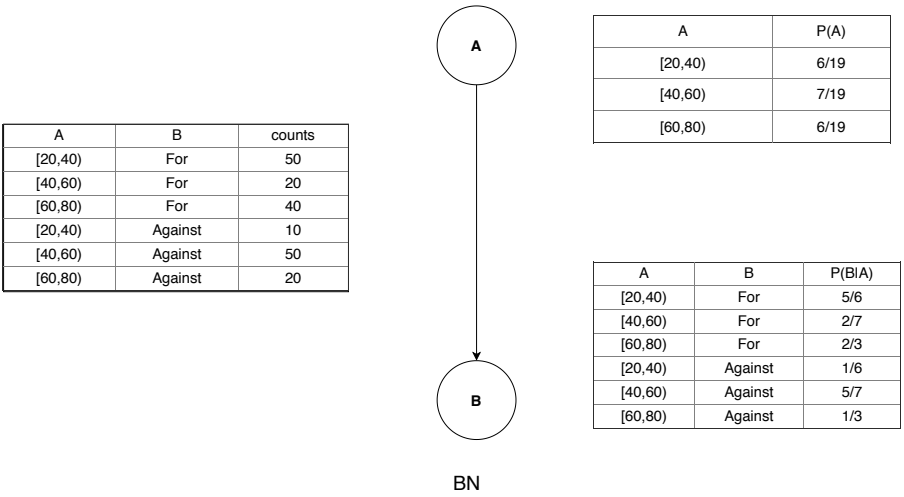
Table 14 Aggregate-level dataset: sample data of TES and VC. TES(left) contains an aggregated count of people's age and attitudes toward bike lanes. VC(right) has the same attributes in individual records

Age	Bike Lanes	counts
[20,40)	For	50
[40,60)	For	20
[60,80)	For	40
[20,40)	Against	10
[40,60)	Against	50
[60,80)	Against	20

Age	Bike Lanes
[20,40)	For
[40,60)	Against
[40,60)	Against
[60,80)	For
...	...

After building a Bayesian network for TES, the join can be performed with the method stated in subsubsection 5.4.2.

One should note that Bayesian networks representing aggregate-level datasets must still obey the LR-learnable constraint specified in subsection 5.3.



A	B	counts
[20,40)	For	50
[40,60)	For	20
[60,80)	For	40
[20,40)	Against	10
[40,60)	Against	50
[60,80)	Against	20

A	P(A)
[20,40)	6/19
[40,60)	7/19
[60,80)	6/19

A	B	P(B|A)
[20,40)	For	5/6
[40,60)	For	2/7
[60,80)	For	2/3
[20,40)	Against	1/6
[40,60)	Against	5/7
[60,80)	Against	1/3

Fig. 11 Aggregate-level data: to construct a Bayesian network for aggregate-level data, the schema is specified with expertise. Parameters are learned with the empirical aggregated counts by Maximum Likelihood.

For example, in Table 15, a census may provide two datasets with one containing Location and Age and the other containing Location and Education. Therefore, we can build Bayesian networks with directed arc (Location → Age) and (Location → Education). However, it is impossible to learn (Age→ Education) or (Education → Age) as they do not appear in any dataset together.

Table 15 Aggregate-level dataset limitation example: location and education do not appear in one sub-table together. Therefore, no edge can be learned between them.

Location	Age	counts
Ward 1	[20,40)	50
Ward 1	[40,60)	20
Ward 1	[60,80)	40
Ward 2	[20,40)	10
Ward 2	[40,60)	50
Ward 2	[60,80)	20
...	...	...

Location	Education	counts
Ward 1	High school or under	20
Ward 1	College or technical school	70
Ward 1	Bachelor degree or higher	20
Ward 2	High school or under	40
Ward 2	College or technical school	30
Ward 2	Bachelor degree or higher	10
...	...	...

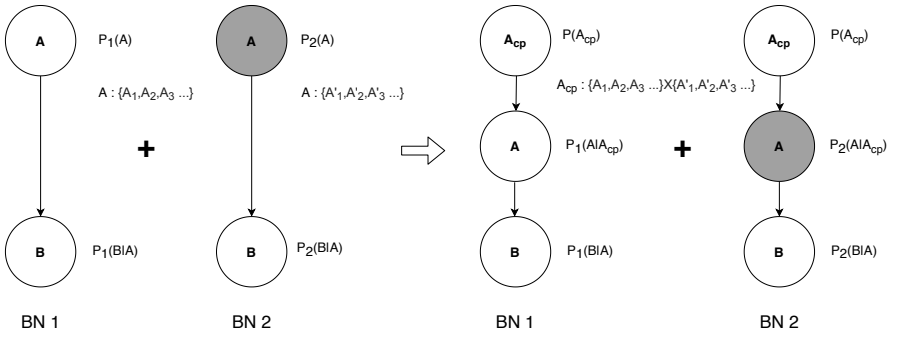
5.4.4 Attribute-domain Mismatch

As illustrated in subsection 4.4, Attribute-domain mismatch refers to when shared attributes are represented differently in two datasets. An example is shown in Figure 12, where the common parent node of BN 1 and BN 2 have differing representations. Node A in the *Ref* model, BN 1 has states: $\{A_1,A_2,A_3\}$ and in BN 2 it has states: $\{A'_1,A'_2,A'_3\}$. When this mismatch occurs, the join can be performed by following three steps.

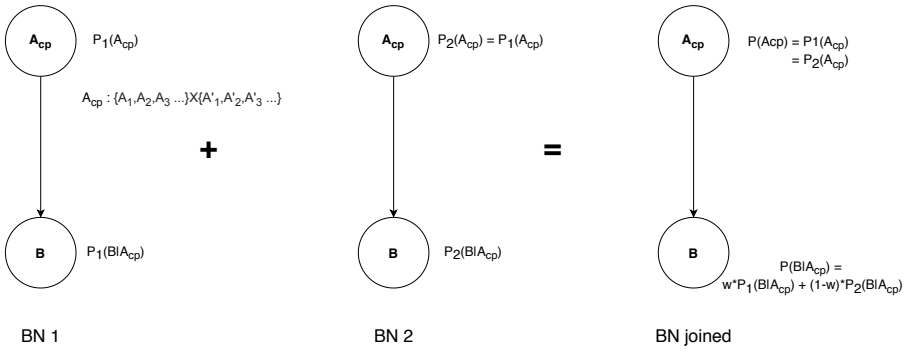
First, as shown in Figure 12(a), take the Cartesian product of the two original state sets to compute a new set of states A_{cp}. This new state set is used to create a new parent node in each local Bayesian network denoted A_cp. Since each state in A_{cp} belongs to a state in each original state set, the conditional distributions of $P_1(A|A_{cp})$ and $P_2(A|A_{cp})$ can be easily learned. When joining the two local Bayesian networks, the prior distribution of $P(A_{cp})$ is selected from the Ref model BN 1, and therefore $P_1(A)$ is redistributed to uniformly when learning $P_1(A_{cp})$.

Second, as demonstrated in Figure 12(b), marginalize out the original parent node A in both networks by using $p(B|A_{cp}) = \sum_A p(B|A) \cdot p(A|A_{cp})$ so that both BN 1 and BN 2 only contain two nodes A_{cp} and B.

Third, join BN 1 and BN 2 using the method described in subsubsection 5.4.2.



(a) Compute the Cartesian product of the two node A state sets to get the states for the new node A_{cp}. Place A_{cp} as the parent node for node A in both BN 1 and BN 2.



(b) Next, marginalize out node A in both BN 1 and BN 2 and perform the join as illustrated in subsubsection 5.4.2.

Fig. 12 Attribute-domain mismatch: BN 1 and BN 2 have the same structure with parent node A and child node B. Node A exhibits attribute-domain since in BN 1, it has states: $\{A_1,A_2,A_3\}$ while in BN 2, it has states $\{A'_1,A'_2,A'_3\}$. The join can be performed using the steps illustrated in Figure 12(a) and Figure 12(b).

As discussed in subsection 4.4, we define three types of Attribute-domain mismatch: Numeric Interval Mismatch, Arbitrary Category Mismatch and, Spatial Representation Mismatch. In this section, we discuss solutions to solve these three types of Attribute-domain mismatch. While they all follow the general process stated in Figure 12, the cross product and prior distribution is calculated slightly differently for each case.

Numeric Interval Mismatch: This type of mismatch is often encountered with datasets containing attributes such as age, income, and height. To handle numerical interval mismatches, we eliminate impossible intervals and redistribute the Ref model's prior distribution according to the interval overlap length between a state in the original state set and a state in the Cartesian product state set. This technique applies both to integer and continuous intervals.

We now demonstrate this technique to resolve the Numeric Interval Mismatch for the Age attribute when joining TES and VC. The age bins that TES uses are {[20,40), [40,60), [60,80)}, whereas VC uses {[20,50), [50,80)}.

Table 16 Numeric Interval Mismatch example with sample data of TES (left) and VC (right). Both datasets have information on Age and Bike Lane Stance, but the age ranges employed in the two surveys differ.

Age	Bike Lanes
[20,40)	For
[20,40)	For
[20,40)	For
[40,60)	Against
[40,60)	Against
[60,80)	For
...	...

Age	Bike Lanes
[20,50)	For
[20,50)	Against
[20,50)	For
[50,80)	Against
[50,80)	Against
[50,80)	Against
...	...

As shown in Figure 13, to represent TES and VC we build models BN 1 and BN 2 where A represents Age and B represents Bike Lanes. Next, we take the Cartesian product of the two-node A state sets to find the new state set for node A_{cp}. This new state set contains only the four valid intervals: [20,40), [40,50), [50,60), and [60,80) since any non-overlapping intervals can be eliminated. For example, [20,40) does not overlap with [50,80).

When calculating the prior distribution of the new node A_{cp}, the distribution of A in the Ref model, BN 1, is selected. Since intervals [40,50] and [50,60] are both in the interval [40,60], the probability of $P_1(A = [40, 60]) = 0.4$ must be distributed to the new states: [40,50] and [50,60]. Without further information, the probability is mapped uniformly according to the length of the interval overlap. In this case, since the two new intervals have equal overlap with the original interval, $P(A_{cp} = [40, 50]) = P(A_{cp} = [50, 60]) = 0.2$. Thus, the CPDs $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ can be derived as shown in Table 17 with invalid intervals being omitted and having zero probability. We can then marginalize out node A from both models and perform the join as described in subsubsection 5.4.4.

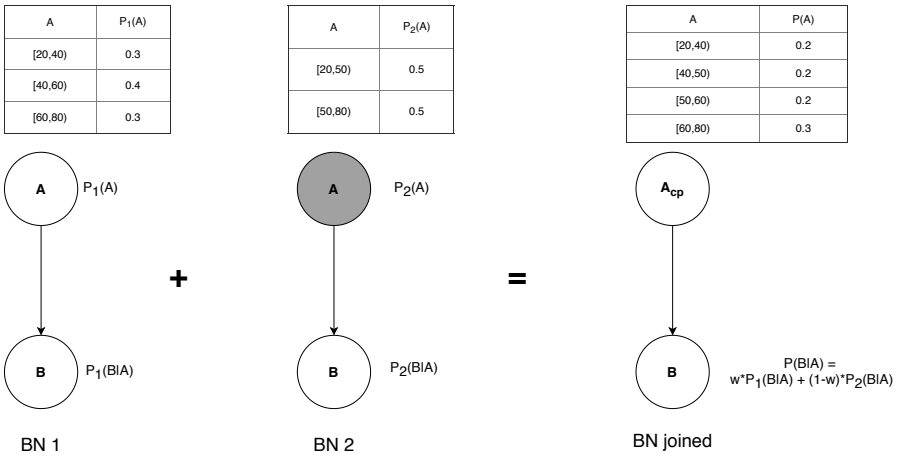


Fig. 13 Numerical Interval Mismatch: A_{cp} has states corresponding to valid intervals obtained from taking the Cartesian product of the differing node A state sets in BN 1 and BN 2. The A_{cp} parameters are based on BN 1's node A parameters, which are re-distributed based on interval length overlap. Node A is marginalized out from BN 1 and BN 2, so the joined Bayesian network, BN-joined, contains only A_{cp} and B.

Table 17 Conditional probability of $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ of numerical mismatch example in Figure 13.

A_{cp}	A (from P_1)	A (from P_2)	$P_1(A_{cp}\|A)$ or $P_2(A_{cp}\|A)$
[20,40)	[20,40)	[20,50)	1
[40,50)	[40,60)	[20,50)	1
[50,60)	[40,60)	[50,80)	1
[60,80)	[60,80)	[50,80)	1

Similarly, the Numeric Interval Mismatch for the Age attribute between census and TES as given in the "2014 Toronto Mayoral Election" use case can be resolved according to the previous procedure.

Arbitrary Category Mismatch For this type of mismatch, the Cartesian product contains all possible state pairs from the two original state sets, unless further information is provided for mappings between state sets. If there exists such a mapping, then we can perform direct mapping. For example, in Table 11, we can assume that the TES Bike Lane responses "Moderately Against" and "Strongly Against" map to the VC response "Against". Additionally, "Strongly For" and "Moderately For" maps to "For". Thus, when joining these two datasets, we can simply map the representations in both datasets to {"For","Neutral","Against"}.

Figure 14 provides an example for performing the join with Arbitrary Category mismatch in the absence of a mapping criteria. In the Ref model, BN 1, node A has three possible states: $\{A_1, A_2, A_3\}$ while in BN 2, it has two possible states: $\{A'_1, A'_2\}$. The state set for A_{cp} is computed by taking the Cartesian

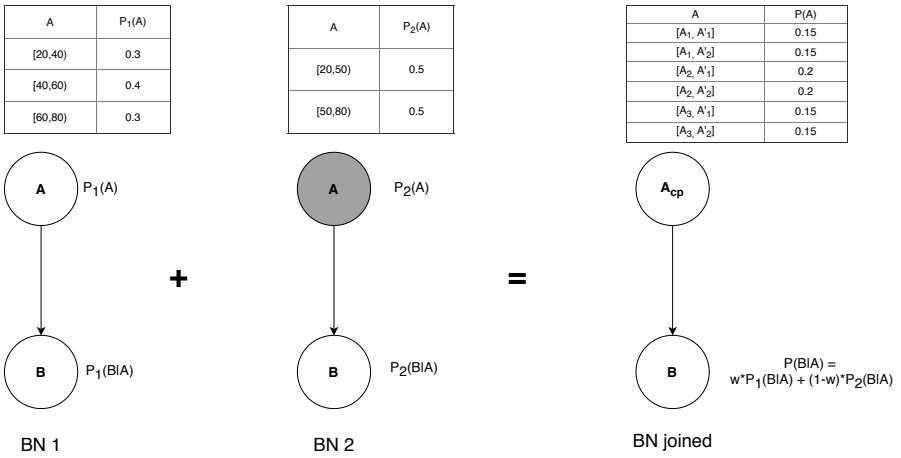


Fig. 14 Arbitrary categories mismatch

Table 18 Conditional probability of $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ for arbitrary categorical mismatch example in Figure 14

A_{cp}	A (from P_1)	A (from P_2)	$P_1(A_{cp}\|A)$ or $P_2(A_{cp}\|A)$
$[A_1, A'_1]$	A_1	A'_1	1
$[A_2, A'_1]$	A_2	A'_1	1
$[A_3, A'_1]$	A_3	A'_1	1
$[A_1, A'_2]$	A_1	A'_2	1
$[A_2, A'_2]$	A_2	A'_2	1
$[A_3, A'_2]$	A_3	A'_2	1

product:

$$\{A_1, A_2, A_3\} \times \{A'_1, A'_2\} = \{[A_1, A'_1], [A_2, A'_1], [A_3, A'_1], [A_1, A'_2], [A_2, A'_2], [A_3, A'_2]\}$$

Without further knowledge, we can not eliminate any of the new states. The prior probability parameters of node A from the *Ref* model are then redistributed uniformly to each corresponding new state. Next, we derive the CPDs $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ as shown in Table 18. We can then perform the marginalization and join as described in subsubsection 5.4.4.

Spatial Representation Mismatch This mismatch occurs commonly in datasets involving locations on Earth's surface expressed as points or polygons. Points are defined by a single x, y coordinate and can represent sampling locations. While polygons consist of 3 or more vertices that are connected and closed and can represent geographic regions. We define two types of Spatial Representation Mismatch as Region-region Mismatch and Region-point Mismatch. The former refers to when one has two overlapping polygon sets that do not cleanly map into each other such as city neighbourhoods and city census tracts. While the latter refers to scenarios where one has a set of point coordinates that can be mapped to a set of polygons.

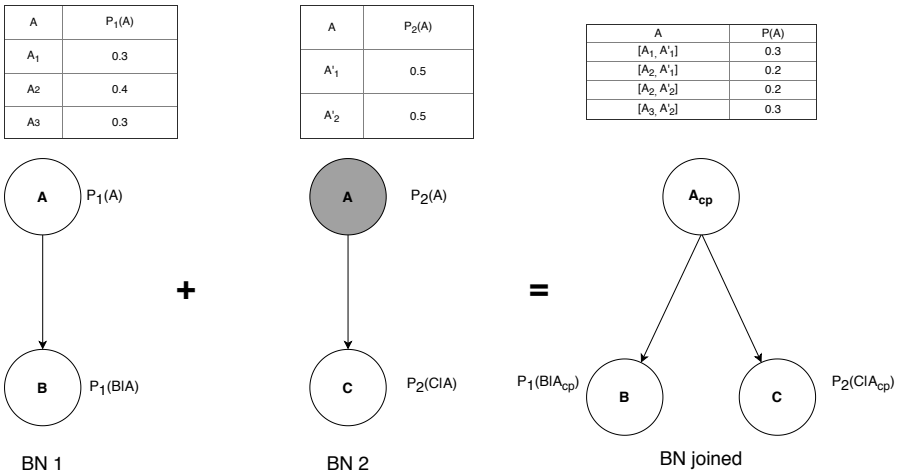


Fig. 15 Spatial mismatch

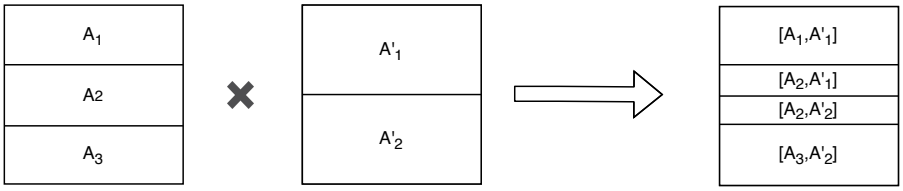


Fig. 16 Spatial cross product to resolve region region mismatch that derives 4 new regions that each map to regions in the original two region sets

Using the simplified example provided in Table 12, we seek to explore the relationship between West Nile virus and median income. Figure 1 demonstrated the Region-region Mismatch between three Voronoi cells constructed from the Chicago West Nile mosquito trap locations and two Chicago community area boundaries.

We proceed to build models BN 1 for the WNV dataset and BN 2 for the median income dataset as shown in Figure 15. Let A represent Region, B represent WNV test results, and C represent median income. Thus, the parent node A has possible values: $\{A_1, A_2, A_3\}$ in BN 1 and $\{A'_1, A'_2\}$ in BN 2. We take the Cartesian product of the two region sets to compute A_{cp}, as shown in Figure 16 to obtain a mutually exclusive and collectively exhaustive region set. This new region set covers the intersection of the two original region sets and eliminates states with non-overlapping regions such as $[A_1, A'_2]$.

In the next step of calculating the new prior distribution A_{cp}, the distribution of A from the Ref model, BN 1, is used as the correct sampling distribution. Since $[A_1, A'_1]$ and $[A_3, A'_2]$ are identical to the regions A_1 and A_3, then $P(A_{cp} = [A_1, A'_1]) = P_1(A = A_1)$ and $P(A_{cp} = [A_3, A'_2]) = P_1(A = A_3)$. Next, the probability $P_1(A = A_1) = 0.4$ is uniformly distributed according to the area overlap of A_1 with new regions $[A_2, A'_1]$ and $[A_2, A'_2]$. Since A_1 can be

mapped exhaustively to just these two new regions and the areas of overlap are equal, $P([A_2, A_1^{'}]) = P([A_2, A_2^{'}]) = \frac{1}{2}P_1(A = A_2) = 0.2$.

Then, the conditional probability of $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ can be derived as shown in Table 19, with entries not shown having probability 0. Although $P_2(A)$ is discarded, it should be noted that $P_2(A)$ having identical values for its two Regions is not because of the equal area of $A_1^{'}$ and $A_2^{'}$ but rather because there are an equal number of observations (just 1) for each community area regarding median income.

Table 19 Conditional probability of $P_1(A_{cp}|A)$ and $P_2(A_{cp}|A)$ of spatial mismatch example in Figure 15

A_{cp}	A (from P_1)	A (from P_2)	$P_1(A_{cp}\|A)$ or $P_2(A_{cp}\|A)$
$[A_1, A_1^{'}]$	A_1	$A_1^{'}$	1
$[A_2, A_1^{'}]$	A_2	$A_1^{'}$	1
$[A_2, A_2^{'}]$	A_2	$A_2^{'}$	1
$[A_3, A_1^{'}]$	A_3	$A_2^{'}$	1

We then can perform the marginalization subsubsection 5.4.4 and the join as shown in subsubsection 5.4.1 .

The same underlying technique also applies to Region-point mismatch where the spatial node in the *Ref* model has regions in its state set, and the spatial node in the *Sec* model has points in its state set. In this case, the Cartesian product of the region set and the point set is simply the region set since points map cleanly into regions. However, mismatch that can not be handled is if the *Ref* model's spatial node state set used points and the *Sec* model used regions since there is no way for regions to be mapped to points.

5.4.5 Performing a query after mismatch handling

After handling mismatches, we can still perform queries when evidence is given as a original state. This requires a rewriting of the original query into multiple queries involving the Cartesian product state set, and aggregating the results.

For example in Figure 13, if one wanted to find the bike lance stance of citizens in both TES and VC, aged 40 to 60 then one can perform the corresponding query $P(B|A = [40, 60))$ on the joined Bayesian network, BN-joined. To perform the query, first rewrite the original query as follows:

$$P(B|A = [40, 60)) = \frac{P(B|A_{cp} = [40, 50)) \cdot P(A_{cp} = [40, 50)) + P(B|A_{cp} = [50, 60)) \cdot P(A_{cp} = [50, 60))}{p(A_{cp} = [40, 50)) + p(A_{cp} = [50, 60))} \quad (2)$$

Next, perform the four queries: $P(B|A_{cp} = [40, 50))$, $P(B|A_{cp} = [50, 60))$, $P(A_{cp} = [40, 50))$ $P(A_{cp} = [50, 60))$ and compute the query result according to Equation 2.

6 Experimental Results

In this section, we use the six datasets described in subsection 3.1 and use cases (1)–(3) from subsection 3.2 as the basis for our experiments, none of which could be addressed by [27] alone. Use case (4) is not explored further since [27] has already experimented with a similar use case using the Bayesian network motif described in subsubsection 5.4.1.

Our aim is to compute a conditional probability query as specified in each use case (1)–(3) described below (nb. all probabilities are indicated in percentage % form to assist interpretation) based on two or more heterogeneous data sources. For purposes of comparison, we provide the same probabilities calculated via different (baseline) methods: (a) a single dataset (which does not require *ppandas*), (b) a *Naive* concatenation merge of two datasets when they share the same variables (which naively ignores any statistical mismatch in the datasets), (c) *ppandas* using two or more datasets (where the datasets are shown with a + separating them), and (d) the actual reference outcome, which is only available for the Toronto Mayoral Election in use case (1). We remark that not all of (a)-(d) are possible in each use case and hence only relevant baselines and reference values are provided when available. Our overall aim in each use case is to discuss the implications of observed differences in the calculations arising from methods (a)-(d).

A toolkit for the automatic application of Bayesian network motifs for unlinked data as described in section 5 is provided in the open source **ppandas** Python package[2]. Furthermore, all of the experimental analysis in this section is provided in reproducible Jupyter notebooks in an *experiments* subfolder of the **ppandas** repository[3].

6.1 Experiment (1): 2014 Toronto Mayoral Election

This experiment leverages the Toronto Election Study (TES) and the 2011 Toronto Census. As illustrated in Figure 17, the model demonstrates the relationship between two common demographic variables and the voting outcome variable:

- Age group (A)
- City ward (W)
- Voting outcome (V)

Since the city's census provides a much more accurate representation of the voting population's demographics, it is used as the Ref distribution for the parent nodes A and W. Therefore, the framework reads in the census population counts for each age group and city ward to learn empirical distributions for A and W using the concepts described in subsubsection 5.4.3. It also uses the concepts described in subsubsection 5.4.4 to handle the numeric interval

[2] https://github.com/D3Mlab/ppandas
[3] https://github.com/D3Mlab/ppandas/tree/master/experiments

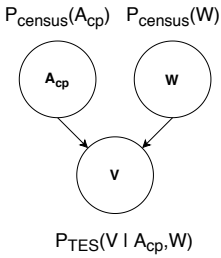


Fig. 17 Experiment (1) model using ppandas to merge TES and census data

mismatch with A_{cp} denoting the Age variable having states derived from taking the Cartesian product between the two different age bin sets employed in the datasets.

Table 20 compares TES's voting distribution without using census data or a Bayesian network versus using ppandas to combine TES and the census then querying $P(V)$. Results for TES (raw data) and *ppandas* using TES+Census are provided in the format "mean percentage (standard deviation)", where the mean and standard deviation are taken w.r.t. five leave-one-out bootstrapped samples of 80% of the data in each respective dataset. We also provide the Actual Outcome of the election results [37] for comparison:

Table 20 Experiment (1): Comparing Voting Outcome Distributions (%)

Candidate	**TES**	***ppandas* TES+Census**	**Actual Outcome**	**TES Error**	***ppandas* Error**
John Tory	48.79 (0.75)	43.12 (1.10)	40.28	8.51	**2.84**
Doug Ford	26.46 (0.55)	28.88 (0.31)	33.73	7.27	**4.85**
Olivia Chow	22.67 (1.03)	24.46 (0.93)	23.15	**0.48**	1.31
Other	2.07 (0.22)	3.53 (0.39)	2.84	0.77	**0.69**

It is evident that much more accurate voting outcome predictions are achieved for the two most popular candidates and that using *ppandas* nearly matches TES's distribution for the third most popular candidate. The small standard deviations show high stability of these results across all replicates of the experimental analysis.

6.2 Experiment (2): Toronto Bike Lanes

This experiment utilizes the two concurrent surveys TES and VC as well as the 2011 Toronto census. The model given in Figure 18 demonstrates the relationship between 4 common demographic attributes and the bike lane stance attribute:

- Age group (A)
- City ward (W)
- Employment status (E)
- Birthplace (B)

- Bike lane stance (L)

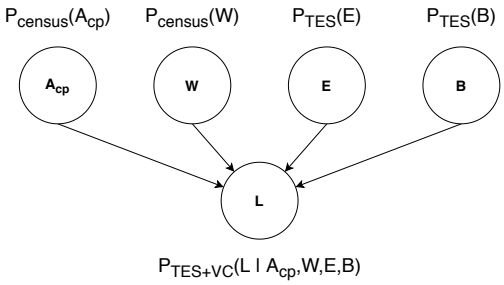


Fig. 18 Experiment (2) model using ppandas to merge TES, VC, and census data

First, the CPD $P_{TES+VC}(L|A, W, E, B)$ was learned from the weighted ratio concept described in subsubsection 5.4.2 that merges TES and VC. Then as in the previous example, numeric interval mismatch for the Age variable was handled when merging census data with TES and VC, at which point A was marginalized out, leaving A_{cp} to remain.

Thus the census' marginal distributions for A and W were used once again but since the census does not contain the demographic attributes E and B, the TES is used as the Ref distribution for these other two attributes.

Table 21 below compares the various results from TES alone, VC alone, and naively stacking TES and VC without a Bayesian network model or the Census. It also shows the results of using ppandas to merge TES and VC without the census ($P_{TES}(A)$ and $P_{TES}(W)$ are used instead) and using ppandas to merge all three data sources to answer the query $P(L)$. All results are provided in the format "mean percentage (standard deviation)", where the mean and standard deviation are taken w.r.t. five leave-one-out bootstrapped samples of 80% of the data in each respective dataset. There is no actual outcome/ground truth available since, unlike the previous experiment which had the actual election outcome, there was no public vote on bike lanes.

Table 21 Experiment (2): Comparing Bike Lane Stance Distributions (%)

Bike Lane Stance	TES	VC	Naive TES+VC	ppandas TES+VC	ppandas TES+VC +Census
For	27.05 (0.38)	79.45 (0.17)	75.27 (0.13)	58.85 (0.23)	51.07 (0.43)
Against	53.73 (0.27)	6.28 (0.06)	10.07 (0.04)	18.96 (0.22)	23.10 (0.32)
Not Sure	19.22 (0.41)	14.27 (0.12)	14.67 (0.10)	22.19 (0.14)	25.83 (0.19)

Clearly, there is a stark difference between TES and VC alone, with over half of TES respondents being against bikes lanes while nearly 80% of VC respondents support them. Merging TES and VC by naively stacking them causes VCs to dominate due to VC containing 10X more samples, and thus the

distribution of variable L is much more similar to VC alone than TES alone. However, when using ppandas, the distribution of L differs with the support for Bike Lanes dropping to less than 60%. The small standard deviations show high stability of these results across all replicates of the experimental analysis.

6.3 Experiment (3): Chicago West Nile

This experiment utilizes the Chicago West Nile Virus Surveillance dataset, Chicago's Community Data Snapshots, as well as the shapefile containing Chicago's community area boundaries. The final model presented in Figure 19 shows the three independent attributes and the two dependent attributes:

- Month (M)
- Number of mosquitoes in trap (N)
- Region (R)
- West Nile virus present (V)
- Median income quintile (I)

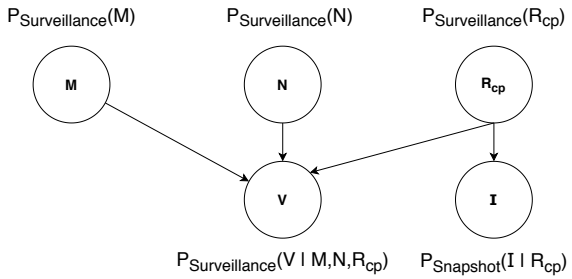


Fig. 19 Experiment (3): West Nile Model

M, N, V are exclusive to the Surveillance dataset and I is provided only in the Snapshots dataset. Thus, the single variable the two datasets share is R. While one could perform a join that handles Region-point mismatch by mapping mosquito point coordinates to community areas, this is unfavourable for two reasons. One, the Ref distribution for R would have to be the Snapshots marginal distribution, which has equal values for all entries since there is only a single median income observation for each community area. Second, it assumes that mosquitoes from each trap stay strictly within a community area.

Instead, an alternative approach is to construct a Voronoi diagram from the mosquito trap locations and obtain another region set composed of the Voronoi cells. With two sets of regions, as shown in Figure 20 and Figure 21, a join that handles Region-region spatial mismatch can be performed using the concepts described in subsubsection 5.4.4. Thus, the Cartesian product of the two region sets yields R_{cp} as shown in Figure 22 and the distribution for node R_{cp} is based on the distribution in the Ref model, WNV Surveillance. Finally,

Figure 23 shows how a specific community area is divided into regions within the Cartesian product.

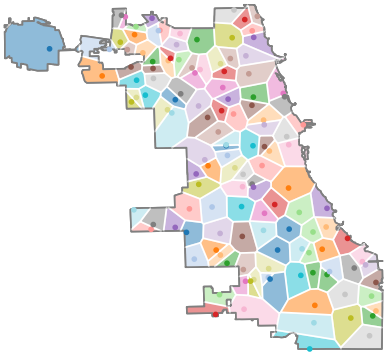


Fig. 20 Experiment (3): Voronoi diagram containing 139 Voronoi cells derived from Chicago mosquito traps

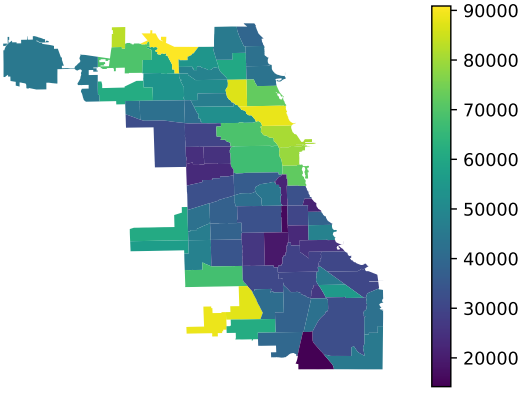


Fig. 21 Experiment (3): Median income for Chicago's 77 Community Areas

Table 22 below provides the results of the query: $P(W|I)$ which is the probability a mosquito will test positive for West Nile for each of the community area quintiles as determined by median income.The highest probability

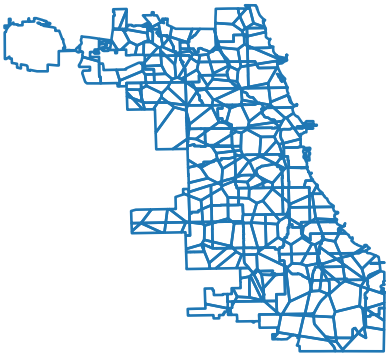


Fig. 22 Experiment (3): Spatial cross product of Chicago community areas and Voronoi diagram cells containing 453 regions

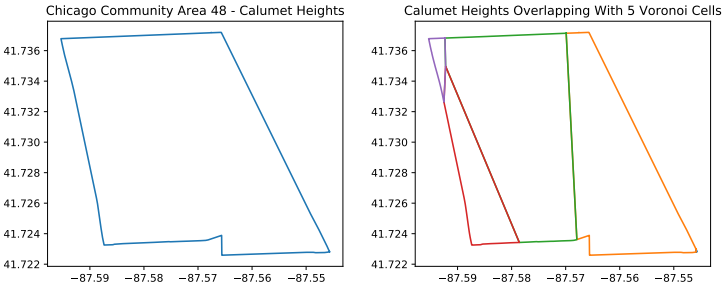


Fig. 23 Experiment (3): The community area, Calumet Heights, is split based on its overlap with 5 Voronoi cells

of testing positive for West Nile came from the community areas in the lowest median income quintile. Additionally, the two highest median income quintiles had higher probabilities than the second and third quintiles.

Table 22 Experiment (3): Median Income and West Nile Virus Positive Tests (%)

Community Area Median Income Quintile	West Nile Virus P(Test Positive)
0 to 0.2	29.12%
0.2 to 0.4	16.60%
0.4 to 0.6	16.71%
0.6 to 0.8	20.92%
0.8 to 1	24.28%

We remark that we did not perform held-out bootstrap analysis for this final use case as in the previous use cases since our current methodology for resolving spatial boundary mismatch does not accommodate missing spatial data that would arise from holding out a percentage of the Census tracts. Hence we only report the results in Table 22 on the full datasets.

7 Discussion

We assessed the proposed framework by carrying out three experiments that involved reasoning jointly over multiple real-world datasets. In doing so, we sought to answer these questions:

1. When the actual ground truth is available, does applying the framework to correct sampling bias provide more accurate inference than reasoning over a single dataset?
2. Do the framework's inference results differ from naïvely stacking multiple datasets?
3. Do the framework's inference results corroborate domain-specific literature results?

For (1), the 2014 Toronto Mayoral Election use case results in Table 20 show that ppandas effectively corrected TES's sampling bias for the independent attributes Age and City Ward. Whereby using the census's marginal distributions for the two demographic attributes, ppandas provides better voting outcome predictions than using TES alone. Intuitively, this result is due to TES being an unrepresentative sample across ten age groups and 44 wards. It has unequal probabilities of inclusion, which favour those who are in younger age groups and live in downtown city wards. Therefore, we use the census to re-weigh under-represented groups in the TES population.

One may question why additional available demographic distributions are not used in the model, such as birthplace and employment status. These are omitted from the Experiment 1 model because the census does not provide marginal distributions for these two attributes. Instead, these two other demographic attributes were collected in the 2011 National Household Survey (NHS). The NHS is a voluntary survey unlike the mandatory census, and thus suffers from non-response bias and is generally regarded as less reliable than the census. Additionally, introducing additional independent attributes would cause data sparsity since only 3,000 TES samples would be available to learn a CPD that contains 31,680 states (10 age groups $\times$ 44 city wards $\times$ 6 employment statuses $\times$ 3 birthplaces $\times$ 4 candidates).

It is also important to note that the VC candidate preferences were not included in Experiment 1. The reason being that after participants provided their candidate preference, the online tool informed participants which candidate best aligned with their beliefs. This additional information may have swayed participants' candidate of choice.

For (2), the motivation to merge TES and VC, as discussed in the Toronto Bike Lanes use case, was due to TES's relatively small sample size, N=3000. As noted previously, one cannot accurately learn the CPD $P(L|A, W, E, B)$, which contains 31,680 unique states with a sample of this small scale. Therefore, to compensate for TES's under-coverage, VC's larger sample size (N=33,000) allows one to learn this CPD more reliably. We also assume that the risk of double counting is insignificant since the size of Toronto's voting population is so large that the chance of a citizen completing both the TES and VC is low.

Since there is no ground truth available (bike lanes are not voted on by the public but by the city council), we cannot strictly prove in this case that ppandas provides more accurate inference than naïvely stacking datasets. We can, however, claim that using ppandas provides an alternative Bike Lance Stance distribution derived from a model that contains the weighted merge of TES and VC's CPD $P(L|A, W, E, B)$ and uses the census and TES marginal distributions for each demographic attribute which is believed to be less biased than VC's marginal distributions.

For (3), the Chicago West Nile use case results reflect prior epidemiology research [38], which indicates there is a relationship between low income areas and West Nile virus due to various factors such as poor home maintenance, antiquated water runoff systems, and less political engagement to request spraying efforts. Additionally, Chicago's high income census tracts with a white, older population have been shown to have high West Nile case rates [39], potentially due to unoccupied homes with neglected swimming pools [40].

Since the West Nile virus data source used in the experiment was drawn from a Kaggle competition, it is worth discussing why this model was not entered in that contest. It is important to recognize that ppandas is not intended for predictive analytics as it does not use extensive feature engineering or seek to minimize prediction error. Instead, it is designed to study causal inference and perform probabilistic queries. Thus it would be inappropriate to direct mosquito spraying action based on ppandas inference. We also felt that utilizing additional datasets not included in Kaggle such as the Chicago Community Snapshots provided us with an unfair advantage. Instead, we go beyond the scope of the Kaggle competition and claim ppandas allows one to study the relationship between median income and West Nile virus test results by handling the spatial mismatch in the Surveillance and Snapshot datasets.

It is additionally worth noting that if every community area did not contain a trap, then constructing a Voronoi diagram would be able to provide an estimate for every community area. Fortunately, every community area does contain a trap but the opportunity to provide more accurate results still remains. We also could have padded missing records when no mosquitoes were found, but did not do so since traps could move around and were not tested every week. Overall, the model demonstrates how median income impacts the probability of a trap testing positive for West Nile.

8 Conclusion

In summary, we provide a systematic framework that allows one to perform probabilistic inference over multiple heterogeneous unlinked datasets without having to join datasets together or impute missing attribute values explicitly. We have shown how each individual dataset can be represented as a Bayesian network, but that attempts to naively merge these models into a joint Bayesian network often yield a variety of data mismatches that, if ignored, yield a high risk of obtaining incorrect inferences. To address these issues, we provide a series of Bayesian network motifs to handle common mismatch types that allow one to construct a single unified Bayesian network representative of the global sampling population. This framework's methodology is built into an open source Python package, ppandas, that allows end users working with multiple heterogeneous unlinked datasets to answer probabilistic queries over all datasets as we demonstrated empirically in a range of real-world use cases.

Declarations

- Funding: The research leading to these results received funding from a University of Toronto Deans Spark Professorship award to SS that partially funded the work of YS, AK, and AO.
- Conflict of interest: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: See next point for links to code and experiments that provide information on how to access all data used in this article.
- Code availability: A toolkit for the automatic application of Bayesian network motifs for unlinked data as described in section 5 is provided in the open source **ppandas** Python package[4]. Furthermore, all of the experimental analysis in this section is provided in reproducible Jupyter notebooks in an *experiments* subfolder of the **ppandas** repository[5].
- Authors' contributions: YS and AK developed the ideas, codebase, and experiments under the supervision of SS with additional supervision from AO. YS, AK and AO wrote the paper with input from SS. DS provided assistance with the TES dataset and experimental analysis.

[4] https://github.com/D3Mlab/ppandas
[5] https://github.com/D3Mlab/ppandas/tree/master/experiments

Appendix A Supplemental Material

A.1 Formal Proof

In this section, we provide a mathematical proof for the weighted average approach in subsubsection 5.4.1.

Problem Setup: Given two datasets $D_1(A, B)$, $D_2(A, B)$, with number of samples N_1 and N_2. Two Bayesian Networks with same schema ($A \rightarrow B$) were created from these two tables, denoted as BN_1 and BN_2 respectively. However, $D_1(A, B)$ and $D_2(A, B)$ are drawn from different distribution and the marginals over A are different. Our goal is to learn the parameter of the joined global Bayesian Network BN_{joined} with the same schema ($A \rightarrow B$), where BN_{joined} are drawn from the same population as Dataset 1. How can we learn the parameter of BN_{joined}, denoted as $r(A), r(B|A)$, given BN_1, BN_2 and number of samples N_1, N_2?

In this section, we use $p(A), p(B|A), q(A)$ *and* $q(B|A)$ represents the true distribution of the Bayesian Networks and use $\hat{p}(A), \hat{p}(B|A), \hat{q}(A)$ *and* $\hat{q}(B|A)$ to denote the parameters learned with Maximum Likelihood from the D_1 and D_2. Since the only data available for $p(A)$, $q(A)$ is from D_1, D_2, respectively. We can claim that $\hat{p}(A), \hat{q}(A)$ can be used as approximation for the true distribution $p(A)$ and $q(A)$.

Solution: The conditional probability distributions(CPDs) in BN_1 and BN_2 can be estimated by the empirical counts in $D_1(A, B)$ and $D_2(A, B)$. Suppose we learned CPDs as shown below:

- $BN_1 : \hat{p}(A), \hat{q}(B|A)$
- $BN_2 : \hat{q}(A), \hat{q}(B|A)$

Then, BN_{joined} can be learned using distributions as following:

- $A : \hat{p}(A)$
- $A \rightarrow B : \hat{p}(B|A), \hat{q}(B|A)$

The prior distribution of $r(A)$ can be estimated as $\hat{p}(A)$ because the joined Bayesian Network were drawn from same population as D_1. The conditional distribution $r(B|A)$ can be estimated using a weighted average of $\hat{p}(B|A)$ and $\hat{q}(B|A)$ with formula:

$$\begin{aligned} r(B|A) &= \frac{N_1 \cdot \hat{p}(B|A) + N_2 \cdot \hat{q}(B|A)}{N_1 + N_2} \\ &= w \cdot \hat{p}(B|A) + (1 - w) \cdot \hat{q}(B|A) \\ w &= \frac{N_1}{N_1 + N_2} \end{aligned} \tag{A1}$$

To justify Equation A1, we utilize importance sampling.

Importance Sampling Justification: We wish to learn the maximum likelihood parameter θ for the final Bayesian Network BN_{joined} $(A \rightarrow B)$ given N_1 data samples $\langle a_i, b_i \rangle \sim p(A, B)$ and N_2 data samples $\langle a_j, b_j \rangle \sim$

$q(A, B)$. Since BN_1 was sampled from the same distribution as BN_{joined}, $r(A) = p(A) = \hat{p}(A)$. Besides, because the sampling bias is limited to the marginal distribution, we can assume $r(B|A) = p(B|A) = q(B|A)$.

Suppose we had K samples (a_k, b_k) drawn from r, then we can learn the parameters with maximum likelihood:

$$\begin{aligned} \underset{\theta}{\text{argmax}} L(\theta : D) &= \underset{\theta}{\text{argmax}} \prod_{k=1}^{K} r(a_k, b_k : \theta) \\ &= \underset{\theta}{\text{argmax}} \left[\sum_{i=1}^{K} \log \left(r(a_k, b_k : \theta) \right) \right] \\ &= \underset{\theta}{\text{argmax}} \left[\frac{1}{K} \sum_{i=1}^{K} \log \left(r(a_k, b_k : \theta) \right) \right] \end{aligned} \tag{A2}$$

We can view the final expression in Equation A2 as a Monte Carlo estimate of the following equation:

$$\frac{1}{K} \cdot \sum_{k=1}^{K} \log \left(r(a_i, b_i : \theta) \right) = E_{r(a,b)} \left[\log r(a, b : \theta) \right]$$

Next, we can apply importance sampling correction to estimate r with $\langle a_i, b_i \rangle \sim p(a, b)$ and $\langle a_j, b_j \rangle \sim q(a, b)$ by re-weighting these samples:

$$\begin{aligned} E_{r(a,b)} \left[\log r(a, b : \theta) \right] &= E_{p(a,b)} \left[\frac{r(a, b)}{p(a, b)} \log r(a, b : \theta) \right] \\ E_{r(a,b)} \left[\log r(a, b : \theta) \right] &= E_{q(a,b)} \left[\frac{r(a, b)}{q(a, b)} \log r(a, b : \theta) \right] \end{aligned}$$

We can then further combine these two expectations as a weighted average to estimate $E_{r(a,b)}$. Since every record in the two tables is treated equally, the weight factor is assigned as $\lambda_1 = \frac{N_1}{N_1+N_2}, \lambda_2 = \frac{N_2}{N_1+N_2}$.

$$
\begin{aligned}
E_{r(a,b)}\left[\log r(a,b:\theta)\right] &= \lambda_1 \cdot E_{p(a,b)}\left[\frac{r(a,b)}{p(a,b)}\log r(a,b:\theta)\right] \\
&\quad + \lambda_2 \cdot E_{q(a,b)}\left[\frac{r(a,b)}{q(a,b)}\log r(a,b:\theta)\right] \\
&= \frac{N_1}{N_1+N_2}\cdot\frac{1}{N_1}\cdot\sum_{i=1}^{N_1}\frac{r(a_i,b_i)}{p(a_i,b_i)}\log\Big(r(a_i,b_i:\theta)\Big) \\
&\quad + \frac{N_2}{N_1+N_2}\cdot\frac{1}{N_2}\cdot\sum_{j=1}^{N_2}\frac{r(a_i,b_i)}{q(a_i,b_i)}\log\Big(r(a_j,b_j:\theta)\Big) \\
&= \frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\frac{r(b_i|a_i)r(a_i)}{p(b_i|a_i)p(a_i)}\log\Big(r(a_i,b_i:\theta)\Big) \\
&\quad + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(b_j|a_j)r(a_j)}{q(b_j|a_j)p(a_j)}\log\Big(r(a_j,b_j:\theta)\Big)
\end{aligned}
$$

Since $r(B|A) = p(B|A) = q(B|A)$ and $r(A) = p(A)$, we know that $r(b_i|a_i) = p(b_i|_i)$, $r(b_j|a_j) = q(b_j|a_j)$,and $r(a_i) = p(a_i)$. Plugging them back into the equation, we get:

$$
\begin{aligned}
&= \frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(a_i,b_i:\theta)\Big) + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(a_j,b_j:\theta)\Big) \\
&= \frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(a_i,b_i:\theta)\Big) + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(a_j,b_j:\theta)\Big) \\
&= \frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(a_i:\theta)r(b_i|a_i:\theta)\Big) + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(a_j:\theta)r(b_j|a_j:\theta)\Big) \\
&= \underbrace{\frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(a_i:\theta)\Big) + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(a_j:\theta)\Big)}_{(7.1)} \\
&+ \underbrace{\frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(b_i|a_i:\theta)\Big) + \frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(b_j|a_j:\theta)\Big)}_{(7.2)}
\end{aligned}
$$

The likelihood decomposes into two separate terms: (7.1) and (7.2). Equation (7.1) is associated with only $r(a)$ and (7.2) with only $r(b|a)$. Therefore, (7.1) can be used for estimating $r(a)$ and (7.2) can be used for estimating $r(b|a)$. Because these terms involve disjoint parameters, they can be optimized separately.

To compute $r(B|A)$, we perform maximization over term (7.2):

$$\frac{1}{N_1+N_2}\sum_{i=1}^{N_1}\log\Big(r(b_i|a_i:\theta)\Big)+\frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\frac{r(a_j)}{q(a_j)}\log\Big(r(b_j|a_j:\theta)\Big)$$

Without loss of generality, we assume that B and A are two Bernoulli random variables. Let $\theta_{B|a_0}(\theta_{B|a_1})$ denote the probability of $b_i=1$ given that $a_i=0(a_i=1)$. Therefore, Equation A.1 can be decomposed as following:

$$\underbrace{\frac{1}{N_1+N_2}\sum_{i|\{a_i=0\}}\log\Big(r(b_i|a_i;\ \theta)\Big)+\frac{1}{N_1+N_2}\sum_{j|\{a_j=0\}}\frac{r(a_j)}{q(a_j)}\log\Big(r(b_j|a_j;\ \theta)\Big)}_{(7.3)}$$

$$+\underbrace{\frac{1}{N_1+N_2}\sum_{i|\{a_i=1\}}\log\Big(r(b_i|a_i;\ \theta)\Big)+\frac{1}{N_1+N_2}\sum_{j|\{a_j=1\}}\frac{r(a_j)}{q(a_j)}\log\Big(r(b_j|a_j;\ \theta)\Big)}_{(7.4)}$$

Now, we focus on maximize the term (7.3) to find the optimal value for $\theta_{B|a_0}$. The maximization on term(7.4) is identical and independent.

$$\begin{aligned}
&\frac{1}{N_1+N_2}\sum_{i|\{a_i=0\}}\log\Big(r(b_i|a_i;\ \theta)\Big)+\frac{1}{N_1+N_2}\sum_{j|\{a_j=0\}}\frac{r(a_j)}{q(a_j)}\log\Big(r(b_j|a_j;\ \theta)\Big)\\
=&\frac{1}{N_1+N_2}\sum_{i|\{b_i=1,a_i=0\}}\log(\theta_{B|a_0})+\frac{1}{N_1+N_2}\sum_{i|\{b_i=0,a_i=0\}}\log(1-\theta_{B|a_0})\\
+&\frac{1}{N_1+N_2}\sum_{j|\{b_j=1,a_j=0\}}\frac{r(a_j)}{q(a_j)}\log(\theta_{B|a_0})+\frac{1}{N_1+N_2}\sum_{i|\{b_j=0,a_j=0\}}\frac{r(a_j)}{q(a_j)}\log(1-\theta_{B|a_0})
\end{aligned}$$

Letting $\#[\cdot]$ denote the count of data i meeting the specified criteria of its argument $\cdot$, then:

$$\begin{aligned}
=&\frac{1}{N_1+N_2}\cdot\Bigg(\#[b_i=1,a_i=0]\log(\theta_{B|a_0})+\#[b_i=0,a_i=0]\log(1-\theta_{B|a_0})\\
+&\#[b_j=1,a_j=0]\frac{r(a=0)}{q(a=0)}\log(\theta_{B|a_0})+\#[b_j=0,a_j=0]\frac{r(a=0)}{q(a=0)}\log(1-\theta_{B|a_0})\Bigg)\\
=&\frac{1}{N_1+N_2}\cdot\Bigg[\Bigg(\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]*\frac{r(a=0)}{q(a=0)}\Bigg)\log(\theta_{B|a_0})\\
+&\Bigg(\#[b_i=0,a_i=0]+\#[b_j=0,a_j=0]*\frac{r(a=0)}{q(a=0)}\Bigg)\log(1-\theta_{B|a_0})\Bigg]
\end{aligned}\tag{A3}$$

To maximize Equation A3, we can take derivative w.r.t. $\theta_{B|a_0}$ and set the value to 0. Since $r(a_0)=p(a_0)=\hat{p}(a_0)$, $q(a_0)=\hat{q}(a_0)$, we get:

$$
\begin{aligned}
0 &= \frac{1}{N_1+N_2}\cdot\left[\frac{\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}}{\theta_{B|a_0}}\right.\\
&\left.-\frac{\#[b_i=0,a_i=0]+\#[b_j=0,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}}{1-\theta_{B|a_0}}\right]\\
&\Rightarrow \theta_{B|a_0}\\
&= \frac{\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}}{\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}+\#[b_i=0,a_i=0]+\#[b_j=0,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}}\\
&= \frac{\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]\cdot\frac{\hat{p}(a=0)}{\hat{q}(a=0)}}{\#[a_i=0]+\#[a_j=0]\cdot\frac{\hat{p}(a)=0}{(a)=0}}\\
&= \frac{1/\#[a_i=0]}{1/\#[a_i=0]}\cdot\frac{\#[b_i=1,a_i=0]+\#[b_j=1,a_j=0]\cdot\frac{\#[a_i=0]}{N_1}\cdot\frac{N_2}{\#[a_j=0]}}{\#[a_i=0]+\#[a_j=0]\cdot\frac{\#[a_i=0]}{N_1}\cdot\frac{N_2}{\#[a_j=0]}}\\
&= \frac{\hat{p}(b_i=1|a_i=0)+\hat{q}(b_j=1|a_j=0)\cdot\frac{N_2}{N_1}}{1+\frac{N_2}{N_1}}\cdot\frac{N_1}{N_1}\\
&= \frac{N_1}{N_1+N_2}\cdot\hat{p}(b_i=1|a_i=0)+\frac{N_2}{N_1+N_2}\cdot\hat{q}(b_j=1|a_j=0)\\
&= \boxed{w\cdot\hat{p}(b_i=1|a_i=0)+(1-w)\cdot\hat{q}(b_j=1|a_j=0);\quad w=\frac{N_1}{N_1+N_2}}
\end{aligned}
$$

Therefore, the maximum likelihood parameters for edge $A \rightarrow B$ can be learned as a weighted average of the empirical conditional probabilities of data samples from the two tables.

References

[1] McGregor, M., Moore, A., Stephenson, L.: What Is Aggregate-level Data? https://www.cihi.ca/en/faq/what-is-aggregate-level-data

[2] Levy, A., Rajaraman, A., Ordille, J.: Querying heterogeneous information sources using source descriptions. VLDB (1996)

[3] Duschka, O., Genesereth, M.: Answering recursive queries using views. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems - PODS (1997). https://doi.org/10.1145/263661.263674

[4] Qian, X.: Query folding, pp. 48–55 (1996). https://doi.org/10.1109/ICDE.1996.492088

[5] Pottinger, R., Halevy, A.: Minicon: A scalable algorithm for answering queries using views. VLDB Journal **10** (2001). https://doi.org/10.1007/s007780100048

[6] Koller, D.: Probabilistic relational models, pp. 3–13 (1999). https://doi.org/10.1007/3-540-48751-4_1

[7] Getoor, L., Mihalkova, L.: Learning statistical models from relational data, pp. 1195–1198 (2011). https://doi.org/10.1145/1989323.1989451

[8] Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press, ??? (2007)

[9] Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases vol. 3, (2011). https://doi.org/10.2200/S00362ED1V01Y201105DTM016

[10] den Broeck, G.V., Suciu, D.: Query processing on probabilistic data: A survey. Foundations and Trends® in Databases **7**(3-4), 197–341 (2017). https://doi.org/10.1561/1900000052

[11] Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association **64**(328), 1183–1210 (1969) https://arxiv.org/abs/https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1969.10501049. https://doi.org/10.1080/01621459.1969.10501049

[12] Christen, P.: Data Matching, (2012). https://doi.org/10.1007/978-3-642-31164-2

[13] Harron, K., Goldstein, H., Dibben, C.: Methodological developments in data linkage (2015)

[14] Steorts, R., Hall, R., Fienberg, S.: Smered: A bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Association **111** (2014). https://doi.org/10.1080/01621459.2015.1105807

[15] Winkler, W.: Matching and record linkage. Wiley Interdisciplinary Reviews: Computational Statistics **6** (2014). https://doi.org/10.1002/wics.1317

[16] Rässler, S.: Statistical matching. a frequentist theory, practical applications, and alternative bayesian approaches **168** (2002)

[17] Kim, J.-k., Rao, J.: Combining data from two independent surveys: A model-assisted approach. Biometrika **99** (2012). https://doi.org/10.2307/41720674

[18] Durrant, G.: Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. International Journal of Social Research Methodology - INT J SOC RES METHODOL **12**, 293–304 (2009). https://doi.org/10.1080/13645570802394003

[19] Carpenter, J., Kenward, M.: Multiple imputation and its application (2012). https://doi.org/10.1002/9781119942283

[20] Metcalf, P., Scott, A.: Using multiple frames in health surveys. Statistics in medicine **28**, 1512–23 (2009). https://doi.org/10.1002/sim.3566

[21] Lohr, S., Raghunathan, T.: Combining survey data with other data sources. Statistical Science **32**, 293–312 (2017). https://doi.org/10.1214/16-STS584

[22] Lohr, S., Rao, J.N.K.: Estimation in multiple-frame surveys. Journal of the American Statistical Association **101**, 1019–1030 (2006). https://doi.org/10.2307/27590779

[23] Jin, S., Komaragiri, V., Rahman, T., Gogate, V.: Learning tractable probabilistic models from inconsistent local estimates. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 10367–10379. Curran Associates, Inc., ??? (2022)

[24] Azzimonti, L., Corani, G., Zaffalon, M.: Hierarchical estimation of parameters in bayesian networks. Computational Statistics and Data Analysis **137**, 67–91 (2019). https://doi.org/10.1016/j.csda.2019.02.004

[25] Azzimonti, L., Corani, G., Scutari, M.: Structure Learning from Related Data Sets with a Hierarchical Bayesian Score. In: Jaeger, M., Nielsen, T.D. (eds.) Proceedings of the 10th International Conference on Probabilistic Graphical Models. Proceedings of Machine Learning Research, vol. 138, pp. 5–16. PMLR, ??? (2020)

[26] Mian, O., Kamp, M., Vreeken, J.: Information-theoretic causal discovery and intervention detection over multiple environments. Proceedings of the AAAI Conference on Artificial Intelligence **37**(8), 9171–9179 (2023). https://doi.org/10.1609/aaai.v37i8.26100

[27] Zhang, B., Sanner, S., Bouadjenek, M.R., Gupta, S.: Bayesian networks for data integration in the absence of foreign keys. IEEE Transactions on Knowledge and Data Engineering **PP**, 1–1 (2019). https://doi.org/10.1109/TKDE.2019.2940019

[28] McGregor, M., Moore, A., Stephenson, L.: The Toronto Election Study. http://www.torontoelectionstudy.com/data

[29] Company, C.B.: CBC Vote Compass 2014 Toronto. https://www.cbc.ca/news2/interactives/votecompass/toronto2014.html

[30] Planning, T.C.: 2011 Toronto Ward Profiles. https://open.toronto.ca/dataset/ward-profiles-2014-2018-wards/

[31] of Public Health (CDPH), C.D.: West Nile Virus Prediction. https://

www.kaggle.com/c/predict-west-nile-virus/data

[32] Survey, A.C.: March 2014 Chicago Community Data Snapshots. https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data/resource/0873e1e4-5160-4396-a2f6-b3961b88852a?inner_span=True

[33] of Chicago, C.: Chicago Community Area Boundaries. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

[34] Bhattacharya, P., Gavrilova, M.L.: Voronoi diagram in optimal path planning. In: 4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD 2007), pp. 38–47 (2007)

[35] Druzdzel, M., Díez, F.: Combining knowledge from different sources in causal probabilistic models. Journal of Machine Learning Research **4**, 295–316 (2003). https://doi.org/10.1162/153244304773633834

[36] Díez, F., Mira, J., Iturralde, E., Zubillaga, S.: Diaval, a bayesian expert system for echocardiography. Artificial intelligence in medicine **10**, 59–73 (1997). https://doi.org/10.1016/S0933-3657(97)00384-9

[37] Wikipedia: 2014 Toronto Mayoral Election. https://en.wikipedia.org/wiki/2014_Toronto_mayoral_election

[38] Harrigan, R., Thomassen, H., Buermann, W., Cummings, R., Kahn, M., Smith, T.: Economic conditions predict prevalence of west nile virus. PloS one **5**, 15437 (2010). https://doi.org/10.1371/journal.pone.0015437

[39] Ruiz, M., Walker, E., Foster, E., Haramis, L., Kitron, U.: Association of west nile virus illness and urban landscapes in chicago and detroit. International journal of health geographics **6**, 10 (2007). https://doi.org/10.1186/1476-072X-6-10

[40] Chung, W., Buseman, C., Joyner, S., Hughes, S., Fomby, T., Luby, J., Haley, R.: The 2012 west nile encephalitis epidemic in dallas, texas. JAMA: the journal of the American Medical Association **310**, 297–307 (2013). https://doi.org/10.1001/jama.2013.8267