

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### T1 Redesign: T1 Partnership Identification Process

by Shaundon Holmstrom

Release date: September 8, 2025



Statistics  
Canada

Statistique  
Canada

Canada

# T1 Redesign: T1 Partnership Identification Process

Shaundon Holmstrom<sup>1</sup>

## Abstract

In Canada, T1 Tax forms are used to report personal income, whether earned as an employee or through self-employment. Income from self-employment, or “T1 Business Income” is reported by sole proprietorships or partnerships. A T1 partnership involves two or more legal entities jointly filing for a shared business. T1 business data is received as individual filings, meaning partnerships are received separately for each partner. Internal record linkage within the T1 business database is performed to identify partnerships and prevent overcoverage within the final population of T1 businesses. This new T1 partnership identification process takes advantage of newer algorithms, such as DBSCAN numerical clustering fuzzy matching, to identify internal linkages. Graph theory is used to construct the list of partnerships from the row-pairs identified in the linkage process.

Key Words: Record linkage; DBSCAN; Graph theory; Administrative data.

## 1. Introduction

### 1.1 T1 Data

T1 tax forms are used to report personal income, including income earned by self-employed individuals. These businesses are unincorporated and file their taxes using personal income tax forms and schedules. Examples of T1 businesses include rental properties, selling handmade goods from a personal store, or selling produce from personally owned farmland.

While business activities conducted by T1 businesses are not exclusive to unincorporated entities, many business owners choose to remain unincorporated for various reasons, such as greater business flexibility, reduced administrative burdens, and lower filing fees.

### 1.2 Organization of T1 Businesses

T1 businesses are organized as sole proprietorships and partnerships. A sole proprietorship consists of a single filing from a single business owner. T1 businesses can report as partnerships, which is a business consisting of two or more partners declaring a single, shared, unincorporated business.

While partnerships between other legal entities do exist, such as other partnerships, incorporated businesses, or charities, they are considered out-of-scope for the T1 partnership process.

### 1.3 How Partnerships File

Partners within a partnership are instructed to submit a filing containing the total financial activity of the business. Ergo, a partner should not submit what they spent individually on utility expenses, but instead the total cost of all utility expenses for the business during a specified reporting period. Absent of all other information, the financial statements of partners within a partnership should be similar.

---

<sup>1</sup> Shaundon Holmstrom, Statistics Canada, 150 Tunney’s Pasture Driveway, Ottawa ON, Canada, K1A 0T6, shaundon.holmstrom@statcan.gc.ca

Business information, such as Business Name when present, Business Address, Business Postal Code, and North American Industry Classification System (NAICS) should match, or be very similar between individual declarations. Partnerships may also register a Business Number, which would provide a unique key to identify the business.

Partners are also able to declare their percentage share of interest in the business, as well as the names of other partners within the partnership.

## **1.4 The Challenge**

T1 filings are received individually from each business owner. For sole proprietorships, the filing is treated as the complete business declaration. Partnerships, however, are received as multiple individual filings, one from each business owner. T1 legislation does not require filers to submit a unique key to identify their business, and partnership identification is impacted as a result.

A major component of T1 business data processing is to create a final population of T1 businesses for statistical purposes. To do so, internal record linkage is required to create a single business declaration for a partnership from the individual statements filed by each partner. Failing to identify all partners that constitute a partnership would result in overcoverage of T1 business population: duplicates of the partnership would be present in the business population. Incorrectly identifying a sole proprietorship as a partnership would result in undercoverage in the business population: the information of one or more sole proprietorships would be missing in the T1 business population.

## **1.5 Why a Redesign?**

The original T1 partnership identification process went into production more than 15 years ago, as of writing. Since then, the needs of data users have evolved significantly, as have available technologies, necessitating a redesign of the processing system.

Transitioning to a completely new system allowed for the integration of contemporary data science methods, machine learning, and complex data structures integral to the new methodology.

# **2. Linking T1 Records**

## **2.1 Steps of the Linkage Process**

The new linkage process operates in three major steps:

1. Numerical clustering on financial columns.
2. Create a set of pair-wise links within each cluster created in (1.)
  - a. Use string comparison methods to determine matching columns between row-pairs.
  - b. Keep row-pairs that satisfy minimum matching criteria.
3. Create a graph using the set of row-pairs kept from (2.)

## **2.2 Financial Clustering**

As partnerships should, ideally, have similar if not identical values in financial fields, the set of potential links is done via clustering the reported financial fields.

The partnership process uses a subset of financial fields that are frequently reported and with the low variance. The fields were chosen via a combination of subject matter expertise, as well as identifying the financial fields with the highest frequency of zero-variance between partnerships identified in the old Partnership Process, e.g. Total Revenue, Total Expenses, Purchases / Costs of Materials, Telephone and Utilities. Financial fields specific to the individual or

with a high probability of reporting at the personal level rather than the business level were excluded from the process, e.g. Travel Expenses, After-tax Income.

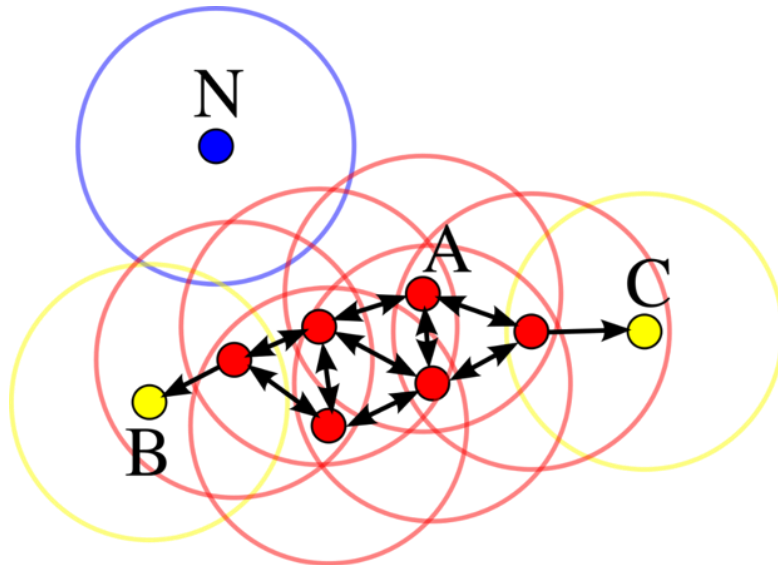
### 2.2.1 The Clustering Algorithm: DBSCAN

DBSCAN is a numerical clustering algorithm designed to find clusters in a large, noisy dataset (Ester et al., 1996). DBSCAN works by calculating a distance metric between two points, and choosing to cluster those points if they are within a specified error bound Epsilon ( $\epsilon$ ) of one another. Any points not within  $\epsilon$  of any other point are not clustered.

DBSCAN can create clusters of arbitrary size and shape and does not require the user to pre-specify the number of clusters beforehand (Ester et al., 1996). The only required parameters given by the user are the distance function and allowed distance between two points ( $\epsilon$ ).

Due to the nature of T1 data, it is impossible to know beforehand exactly how many partnerships will appear in a given year. Not having to pre-specify the number of clusters, and DBSCAN's ability to handle noise which ought not to be linked are two very attractive properties to handle this problem. Noise Points are assumed to be sole proprietorships, absent of any other criteria.

**Figure 2.2.1-1**  
**Illustrated Example of the DBSCAN Algorithm (Chire, 2011)**



A (red): Centre points. Connected to 2 or more other points.  
B, C (yellow): Edge points. Connected to 1 centre point.  
N: Noise points. Not within  $\epsilon$  of any other point.

## 2.3 Comparing Row-Pairs Within a Cluster

After a set of rows have been assigned to a same cluster, all pairwise comparisons are performed within that cluster. String comparison similarity scores are produced for Business Name and Business Address. Other fields such as Business Number, NAICS, and Partner Name information is compared on an exact match.

### 2.3.1 Column Comparisons

Typos and varying reporting formats are found in the Business Address and Business Name fields. String similarity scores are used to account for reporting inconsistencies. It was determined other fields were either too short or too

specific to be worth accounting for typos or other mistakes, and only exact matches were allowed for fields not the name or address.

Jaro-Winkler similarity was chosen as the distance metric for strings since it handles non-matching string lengths and has proven application in record linkage (Winkler, 1990). Its performance on the T1 data was assessed and determined to be acceptable.

### 2.3.2 Minimum Matching Criteria

Once all matching and non-matching columns have been determined, a set of combinations of matching columns are chosen as “minimum matching criteria”. We keep row-pairs that meet at least one of the minimum matching column combinations.

E.g. A match is considered adequate if the cluster analysis, Business Name, and Business Postal Code are found as matching between two rows.

### 2.4 Graphing Kept Row-Pairs

A graph is a data structure that consists of Nodes, and Edges that connect those Nodes. For the purposes of this project, the nodes in the graph represent rows in the dataset, and the edges are the set of accepted row-pairs.

The remaining list of row-pairs that pass the minimum matching criteria are passed to a graphing library (Hagburg et al, 2008). A graph is created from the row-pairs, and subgraphs are extracted. Each set of joined nodes in the output corresponds to a finalized partnership.

## 3. Fabricated Example Walkthrough of the Partnership Process

### 3.1 Clustering Financials

Assume we have a population of the following six (6) T1 filings. In the first step, we ignore all other business information and focus solely on clustering via financial reports. We run DBSCAN with an error bound Epsilon ( $\epsilon$ ) appropriate for the data domain. For illustrative purposes,  $\epsilon = 200$  will be used in the following example.

**Table 3.1-1**

**Example Data: DBSCAN on T1 Financials**

ID Code	Filer Name	Total Revenue	Total Expenses	Cost of Utilities	Leasing Revenues	DBSCAN Cluster
132_1	John Smith	50,000	1,200	0	0	1
132_2	John Smith	50,100	1,150	0	0	1
865_1	Alice Jones	50,100	1,150	0	0	1
611_1	Eliza Rochefort	25,000	1,500	900	25,000	2
376_1	Emilie Rochefort	25,000	1,500	900	25,000	2
978_3	Marie Dupoint	30,000	800	750	30,000	-1

### 3.2 Matching Pairs Within a Cluster

With the clusters created by DBSCAN, the profile of which is shown in Table 3.2-1, the business’ identifying columns are compared.

Intuitively, it is easy for a human to determine that John Smith's second filing (132\_2) and Alice Jones (865\_1) form a partnership. They have reported the same address, but formatted such that a machine will have trouble matching strictly by string comparison. The different Business Names and Business Number can be attributed to typos. Partner Names are also a clear indication of the partnership.

Eliza and Emilie Rochefort are also obvious candidates for a partnership. They have reported the same address, a partner share of 50%, and a matching Industry Code.

**Table 3.2-1**  
**Example Data: Business Information Columns**

ID Code	Filer Name	DBSCAN Cluster	Business Number	Business Name	Business Address	Industry Code	Partner Percent	Partner Name(s)
132_1	John Smith	1			456 Street	333		
132_2	John Smith	1	885	Gold Star Staffing	#3 467 Street	334	50	Jones
865_1	Alice Jones	1	8885	Gold Star Staff	Unit 3, 467 Street, NW	334		John Smith
611_1	Eliza Rochefort	2			123 Sesame	531	50	
376_1	Emilie Rochefort	2			123 Sesame St	531	50	
978_3	Marie Dupoint	-1			1-476 Street	531		

Within each DBSCAN Cluster, the full set of pairwise comparisons is performed on each of these columns. Cluster 1 will have three pairwise comparisons to make, and Cluster 2 will have a single pair-wise comparison. Marie Dupoint (978\_3) has been assigned no cluster (Noise point) and will not be included moving forward.

**Table 3.2-2**  
**Example Data: Row-Pair Column Comparisons**

ID Code 1	ID Code 2	DBSCAN Cluster	Business Number Match	Business Name Match	Business Address Match	Industry Code Match	Sum Partner Percent = 100	Partner Name Match	Accept Match?
132_1	132_2	1	0	0	0.59	0	0	0	No
132_1	865_1	1	0	0	0.63	0	0	1	No
132_2	865_1	1	0	0.97	0.61	1	0	1	Yes
611_1	376_1	2	0	0	0.89	1	1	0	Yes

The pair (132\_1, 132\_2) is immediately discarded since an individual cannot be a partner with themselves.

The pair (132\_1, 865\_1) is also not accepted. While Alice Jones has reported John Smith as a partner, and that John Smith has two filings within the cluster, no other matching criteria being fulfilled suggests that the match is of poor quality. The match is not accepted.

The pair (132\_2, 865\_1) shows an excellent match on the Business Name and reported partners. A matching industry code is also indicative. The match is accepted.

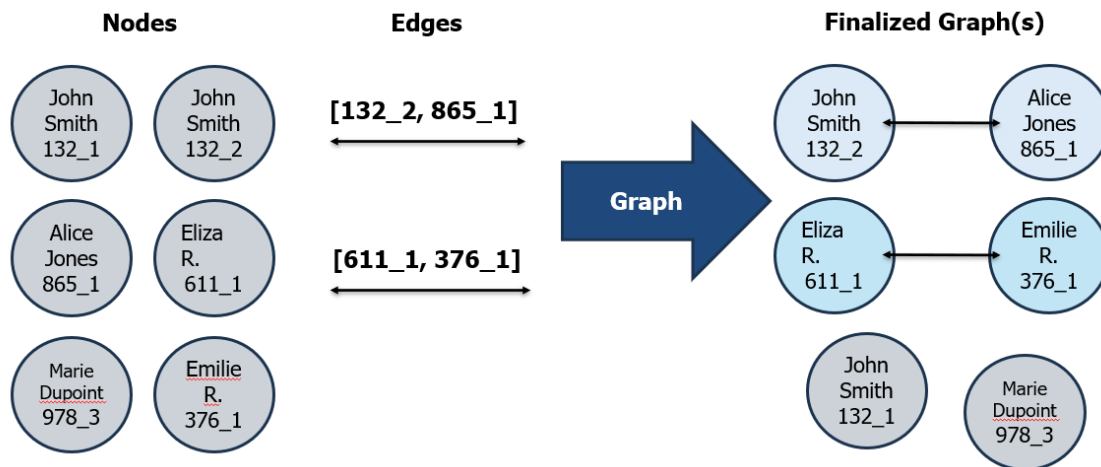
The pair (611\_1, 376\_1) have an adequate match on the Business Address, and Partner Percent that adds up nicely to 100%. The match is accepted.

### 3.3 Graphing Accepted Row-Pairs

Graph theory allows the construction of the final set of partners from the row-pairs identified in (3.2). Inputting the row-pairs into graphing software will create the graph where each resulting subgraph of connected points represents a finalized partnership.

In the provided example, the row-pairs of accepted matches are used as edges within the graph. The output identifies two partnerships of two points each: John Smith (132\_2) and Alice Jones (865\_1); Eliza Rochefort (611\_1) and Emilie Rochefort (376\_1). Two sole proprietorships John Smith (132\_1) and Marie Dupoint (978\_3) are also identified.

**Figure 3.3-1**  
Overview of the Graphing Process



## 4. Conclusion

The redesigned partnership process methodology has produced promising results. Industries with complicated reporting patterns still prove a struggle for the methodology as of writing, but further refinements to the application of DBSCAN and splitting the process by industry are showing improvements to the process. The redesign also means we have a more sophisticated platform, compared to the software we intend to replace, on which we can develop future enhancements.

## References

- Chire, (2011), "DBSCAN-Illustration," *Wikimedia Commons*. Retrieved April 15, 2025 from <https://commons.wikimedia.org/w/index.php?title=File:DBSCAN-Illustration.svg&oldid=872952835>.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231.
- Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008), "Exploring network structure, dynamics, and function using NetworkX", *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pp. 11-16.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida". *Journal of the American Statistical Association*, 84(406), pp. 414-420.

Winkler, W, E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Evaluative Report, Washington DC, United States of America: U.S. Bureau of the Census Stat. Research Div.