Airbnb Listing Networks: Community Detection and Price Influence Across Toronto Neighbourhoods

EECS 4414 Information Networks Final Project Report

Minh Anh Nguyen*
York University
Toronto, Canada
minhanh0@my.yorku.ca

Kirti*
York University
Toronto, Canada
kdhir04@my.yorku.ca

Sourav Chandhok*
York University
Toronto, Canada
sourav@my.yorku.ca

Abstract

This project models Toronto Airbnb listings as an information network in order to study how short-term rental markets cluster across the city and whether network communities help explain listing prices. Each listing is represented as a node, and edges are formed using three sources of information: geographic proximity, shared host ownership, and listing attribute similarity. Three graph variants are compared: a spatial graph, a spatial plus host graph, and a full graph that also includes attribute-similarity edges. Community structure is detected using Louvain and Leiden modularity optimization, and detected partitions are evaluated against official Toronto neighbourhood labels using normalized mutual information and variation of information. The experiments use a cleaned Toronto Airbnb dataset containing 15,809 listings across 140 neighbourhoods. Results show that the full graph becomes connected and produces fewer, larger market communities than the spatial-only graph. Spatial-only communities align most strongly with neighbourhood boundaries, while full-network communities reveal broader market structure. The detected full-graph communities correspond to interpretable market segments: a dominant citywide background market, several tight premium downtown clusters, and a few budget private-room pockets. Adding community membership to a price model gives a small but measurable improvement in predictive accuracy.

1 Introduction

Short-term rental platforms such as Airbnb create dense local marketplaces in which location, host behaviour, listing type, and guest-facing attributes interact. A listing is not isolated: it competes with nearby rentals, may be managed by the same host as other listings, and often resembles other listings with similar room types, capacity, or review profiles. These relationships suggest that Airbnb data can be represented as an information network rather than only as a standard tabular dataset.

This project studies Toronto Airbnb listings using network analysis. The main goal is not simply to build a graph, but to identify data-driven Airbnb market communities and test whether they reveal structure beyond official neighbourhood labels. In this study, a community represents a group of listings that are more strongly related to one another than to the rest of the market through spatial proximity, shared-host ownership, or similar listing attributes. These communities can be interpreted as functional Airbnb market segments. Official neighbourhoods are administrative units, but rental markets may not follow those boundaries exactly. For example, listings in adjacent downtown areas may belong to the same high-price tourism market, while listings within the same official neighbourhood may be separated by different room types or host strategies.

The project builds three graph variants. Graph A contains only spatial proximity edges. Graph B adds shared-host edges. Graph C further adds attribute-similarity edges. Louvain and Leiden are then used to detect communities in each graph. These algorithms are compared using modularity, community-size distributions, and alignment with official Toronto neighbourhood labels. Finally, a ridge regression price model compares a baseline feature set against an expanded model that includes detected community membership.

The contribution of this project is an end-to-end network-based analysis of Toronto Airbnb listings. The results show that adding non-spatial relationships substantially changes the graph structure: Graph C becomes one connected component and produces larger communities that diverge from official neighbourhoods. These communities map onto interpretable market segments rather than administrative units. However, the price-model improvement from community membership is modest, suggesting that detected communities capture some market signal but do not replace listing-level and neighbourhood-level predictors.

2 Problem Definition

Let $G = (V, E)$ be an undirected weighted graph where each node $v_i \in V$ represents one Airbnb listing. Each listing has geographic coordinates, host identity, official neighbourhood, room type, property type, price, and additional numerical attributes. The project investigates three questions:

RQ1: How does adding shared-host and attribute-similarity relationships change the connectivity of the Airbnb listing network?

RQ2: Do detected communities align with official Toronto neighbourhoods, or do they reveal different market structures?

RQ3: Does community membership improve price prediction beyond listing attributes and official neighbourhood labels?

In this project, detected communities are interpreted as data-driven Airbnb market segments. Listings within the same community are expected to be more closely connected through geography, host behaviour, or listing characteristics, and may therefore represent listings that compete in similar parts of the short-term rental market. The purpose of community detection is therefore to test whether the Toronto Airbnb market contains meaningful structural segments beyond official neighbourhood labels.

*All authors contributed equally to this research.

Edges are defined by three relationship types. For spatial proximity, two listings are connected if they are within a 500 meter radius. The spatial contribution is larger for closer listings:

$$S_{ij} = 1 - \frac{d_{ij}}{500}, \quad d_{ij} \leq 500. \tag{1}$$

For shared host ownership, listings owned by the same host are connected using a sparse nearest-neighbour strategy to avoid creating very large host cliques. For attribute similarity, each listing is connected to its five nearest listings using cosine similarity over standardized numerical features and one-hot encoded categorical features. The final edge weight is

$$w_{ij} = 0.60S_{ij} + 0.25H_{ij} + 0.15A_{ij}, \tag{2}$$

where H_{ij} is the shared-host component and A_{ij} is the attribute-similarity component.

Community detection returns a partition $C = \{C_1, C_2, \ldots, C_k\}$ of the listings. The detected partition is evaluated using modularity, normalized mutual information (NMI), and variation of information (VI). Higher modularity indicates stronger community structure. Higher NMI indicates stronger agreement with official neighbourhood labels, while lower VI indicates a smaller information distance between the two partitions.

In summary, the input to the problem is the set of 15,809 listings together with their geographic, host, and attribute information, and the desired outputs are twofold: (i) a partition of the listings into market communities that maximizes weighted modularity, and (ii) a price model whose error is reduced by incorporating the detected community membership as a feature.

3 Related Work

Community detection is a central task in network analysis. Modularity measures the extent to which a partition has more within-community edges than expected under a null model [5]. Louvain is a widely used greedy modularity optimization method that repeatedly moves nodes into communities and aggregates the network to improve modularity [2]. Leiden improves on Louvain by refining communities and providing stronger guarantees about well-connected communities [3]. Because this project compares market clusters, using both algorithms provides a useful robustness check.

Comparing detected communities with external labels is also a common evaluation strategy. NMI measures statistical dependence between two partitions and is often used for community validation [7]. VI treats the difference between clusterings as an information distance [6]. In this project, the external labels are official Toronto neighbourhoods.

The idea of modelling Airbnb itself as a network is supported by prior work. Teubner [4] analyses Airbnb from a network perspective, studying the web of host–guest connections and arguing that the platform is best understood through the relationships between its actors rather than as a collection of independent listings. Our project adopts the same relational view but applies it to listings rather than host–guest interactions, and uses the resulting structure for community detection and price modelling.

Airbnb pricing has often been studied with hedonic and machine learning models. Prior work has shown that price is affected by capacity, property type, location, reviews, and host-related variables [10, 11]. This project builds on that line of work by treating market structure as a network feature. Unlike that prior pricing work, which treats listings as independent observations, we model the relationships between listings explicitly and test whether detected listing communities add predictive value beyond official neighbourhood labels.

Table 1: Cleaned dataset summary.

Statistic	Value
Listings	15,809
Official neighbourhoods	140
Unique hosts	10,326
Median winsorized price	$125
Mean winsorized price	$164.22
Median review score	4.89
Median availability	249 days

4 Methodology

4.1 Dataset and Preprocessing

The experiments use a cleaned Toronto Airbnb listings dataset derived from the Inside Airbnb data [1], using the November 2025 Toronto snapshot. After removing listings with missing coordinates or prices and dropping duplicate listing IDs, the final dataset contains 15,809 listings across 140 official neighbourhoods and 10,326 unique hosts. Prices were winsorized to reduce the effect of extreme outliers, and the target variable for modelling was the log-transformed winsorized nightly price. Table 1 summarizes the cleaned dataset.

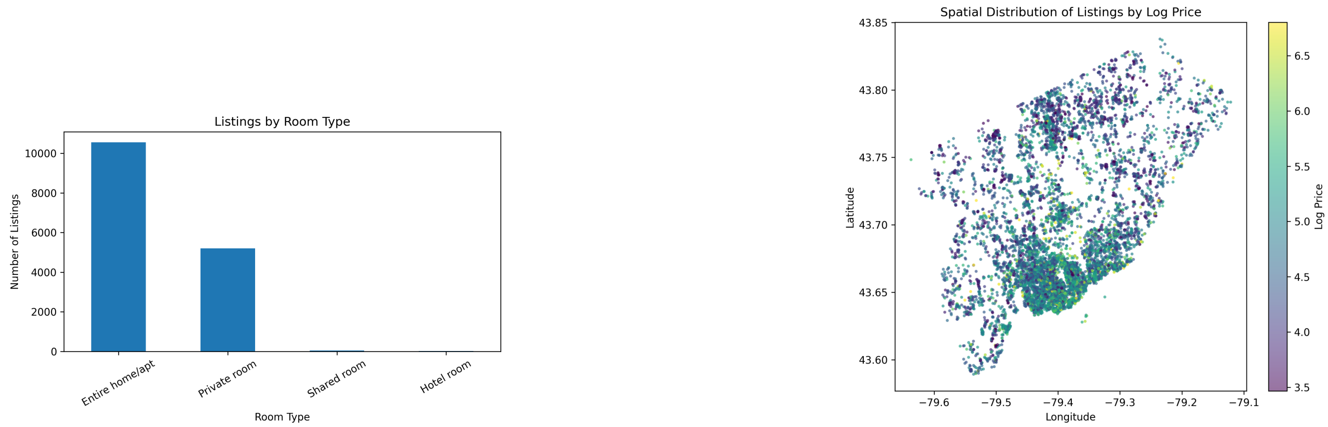
Missing numerical values were replaced with column medians. Missing categorical values were replaced with "Unknown." Numerical variables used in graph construction and price modelling include capacity, bedrooms, beds, availability, number of reviews, review score, and reviews per month. Categorical variables include room type, property type, official neighbourhood, instant booking status, and host superhost status. Figure 1 shows the room-type composition and the spatial price distribution of the cleaned listings, and Figure 2 reports the highest-priced high-listing neighbourhoods that motivate combining location with market structure in the price analysis.

4.2 Graph Construction

Three graph variants were constructed to isolate the effect of adding different relationships:

- **Graph A: Spatial only.** Edges connect listings within 500 meters.
- **Graph B: Spatial + shared host.** Graph A plus sparse same-host edges.
- **Graph C: Spatial + shared host + attribute similarity.** Graph B plus attribute-based nearest-neighbour edges.

Spatial edges were generated efficiently using a BallTree with the haversine distance metric. Shared-host edges were added between each listing and up to five nearest listings owned by the same host. This design avoids forming a complete clique for very large multi-listing hosts, which would dominate the graph and slow computation. Attribute-similarity edges were generated using



(a) Room type distribution.

(b) Spatial distribution coloured by log price.

Figure 1: Exploratory views of the cleaned Toronto Airbnb dataset.

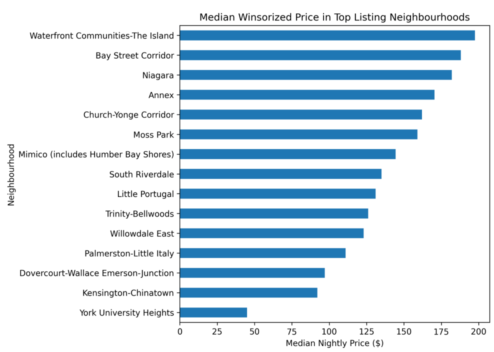


Figure 2: Median winsorized nightly price in the most expensive high-listing Toronto neighbourhoods. This exploratory result motivates using both location and market/community information in the price analysis.

Table 2: Main graph-construction parameters.

Parameter	Value
Spatial radius	500 meters
Shared-host neighbours	5
Attribute neighbours	5
Spatial weight coefficient	0.60
Shared-host coefficient	0.25
Attribute-similarity coefficient	0.15

five nearest neighbours under cosine distance after preprocessing numerical and categorical attributes.

Table 2 summarizes the final parameter settings. These values were selected to keep the graph interpretable and computationally feasible. A dense shared-host clique construction was avoided because a host with many listings would create a large number of edges that would not necessarily represent local market similarity. The sparse nearest-neighbour version preserves the host relationship while limiting this effect.

4.3 Community Detection

Community detection was performed by maximizing modularity. For a weighted graph, modularity is

$$Q = \frac{1}{2m} \sum_{i,j} \left[w_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{3}$$

where w_{ij} is the edge weight between listings i and j, $k_i = \sum_j w_{ij}$ is the weighted degree of node i, m is the total edge weight, c_i is the community of node i, and $\delta(c_i, c_j) = 1$ when the two listings share a community. Intuitively, Q measures how much more edge weight falls inside communities than would be expected if the same edges were rewired at random. A partition with high Q has dense internal connections and sparse connections between communities. Maximizing Q exactly is NP-hard, so heuristic algorithms are used.

Louvain [2] optimizes Q in two repeated phases. In the local-moving phase, each listing is moved to the neighbouring community that yields the largest gain in Q. In the aggregation phase, each community is collapsed into a single super-node, and the two phases repeat on the smaller graph until Q no longer improves. Louvain is fast, but a known weakness is that the communities it returns can be internally disconnected.

Leiden [3] adds a refinement phase between local moving and aggregation: communities are first split into well-connected sub-communities, and aggregation is performed on the refined partition. This guarantees that every detected community is internally connected. In this project Louvain is the main interpretation method and Leiden is a robustness check — if both recover similar structure, the detected market segments are unlikely to be artefacts of one heuristic.

Both algorithms were applied to the weighted graphs, so stronger spatial, host, or attribute relationships contributed more to the optimization. The implementation used NetworkX for graph construction and Louvain [8], and the igraph and leidenalg packages for Leiden. For each run, the number of communities, modularity, largest community size, median community size, and singleton count were recorded.

A known property of modularity maximization is the *resolution limit*: communities smaller than a scale set by the total edge weight of the graph tend to be merged into larger ones. This is directly relevant to Graph C, where attribute-similarity edges act as long-range bridges between spatially distant listings with similar profiles. These bridges raise the modularity reward for merging far-apart clusters, which is why Graph C collapses from roughly 134 communities (Graph A) to only 16, including one dominant community. The dominant community should therefore be read as a broad background market segment rather than a single coherent local cluster.

4.4 Neighbourhood Alignment

Detected communities were compared with official Toronto neighbourhood labels. NMI was used because it gives a normalized agreement score between 0 and 1. VI was also used because it measures the information distance between two partitions. Together, these metrics show whether the detected communities are mostly rediscovering neighbourhood boundaries or capturing broader market relationships.

4.5 Price Influence Analysis

To test whether communities help explain price, two ridge regression models were trained using an 80/20 train-test split. The baseline model used listing attributes and official neighbourhood labels. The expanded model added Graph C Louvain community membership. Ridge regression was selected because the one-hot encoded categorical variables create a high-dimensional feature space, and regularization helps reduce overfitting [9].

Importantly, price is not used anywhere in graph construction: the attribute-similarity edges are built only from capacity, bedrooms, beds, availability, review counts, review score, reviews per month, room type, and property type. Community membership is therefore derived independently of price, so using it as a predictor is not circular and any gain reflects genuine market-structure information rather than leakage of the target.

The prediction target was log-transformed winsorized nightly price. Performance was measured using test R^2, log-scale MAE and RMSE, and approximate dollar-scale MAE and RMSE obtained by exponentiating predicted and actual log prices. Because all reported metrics are computed on a held-out test set, the improvement of the expanded model reflects out-of-sample gain rather than in-sample overfitting.

5 Evaluation and Results

5.1 Graph-Level Results

Table 3 shows how the graph changes as additional relationship types are added. The spatial-only graph has 100 connected components, with the largest component containing 87.8% of listings. Adding sparse shared-host edges reduces the number of components to 53 and increases the largest component to 98.7% of listings. Adding attribute-similarity edges makes Graph C fully connected.

The weighted clustering coefficient decreases from 0.2967 in Graph A to 0.1022 in Graph C. This is expected because attribute-similarity edges connect listings across a broader market space and reduce the purely local nature of the spatial graph. The number of edges grows only moderately from Graph A to Graph C, but the component structure changes substantially, showing that a small number of non-spatial links can strongly affect global connectivity.

5.2 Community Detection Results

Table 4 compares Louvain and Leiden on the three graph variants. Graph A and Graph B have high modularity and many communities. Graph C also has high modularity but produces fewer, larger communities. Louvain finds 16 communities in Graph C, while Leiden finds 18. Their modularity scores are almost identical: 0.7927 for Louvain and 0.7934 for Leiden. The collapse from about 134 communities in Graph A to 16 in Graph C follows directly from the resolution limit discussed in Section 4.3: the five-nearest-neighbour attribute edges link listings across the entire city, so modularity optimization merges spatially separated clusters that share similar listing profiles into a small number of broad market segments.

Figure 3 shows the spatial layout of Graph C communities. Both algorithms identify spatially concentrated clusters in the downtown core and surrounding areas, while also assigning many listings across the city to a large background market community. The Leiden result splits the full graph into slightly more communities, but the overall visual structure remains similar. These communities are examined in more detail, with named market segments, in Section 5.4 (Figure 6).

The community-size plots in Figure 4 show that both algorithms identify a dominant market community. This dominant community contains listings spread across many neighbourhoods, which explains why Graph C has weaker agreement with official neighbourhood labels than Graph A. At the same time, the remaining smaller communities often correspond to dense central areas or distinctive market segments such as private-room clusters.

5.3 Alignment with Official Neighbourhoods

Table 5 reports NMI and VI between detected communities and official neighbourhoods. The spatial-only graph aligns most strongly with official neighbourhoods. Graph A Leiden has the highest NMI at 0.7538 and the lowest VI at 2.7562. As host and attribute edges are added, NMI decreases and VI increases. This indicates that Graph C communities are less similar to official neighbourhood labels.

Figure 5 visualizes the same alignment results. It shows that NMI decreases from Graph A to Graph C, while VI increases. This trend confirms that adding host and attribute relationships moves the detected communities away from official neighbourhood boundaries and toward broader Airbnb market segments.

This result answers RQ2. Spatial proximity alone largely reconstructs neighbourhood-like partitions. The full graph, however, captures relationships that cross administrative boundaries. In other words, official neighbourhoods are useful but incomplete proxies for Airbnb market structure.

5.4 What the Detected Communities Represent

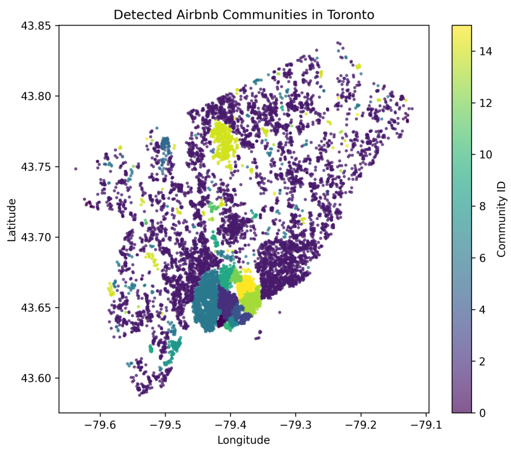
The 16 Graph C Louvain communities are not abstract clusters: each one is an interpretable Airbnb market segment defined by a dominant location, a dominant room type, a characteristic price level, and a geographic footprint. Table 6 lists all 16 communities, Figure 7 summarizes them visually, Figure 6 shows their spatial

Table 3: Graph construction results.

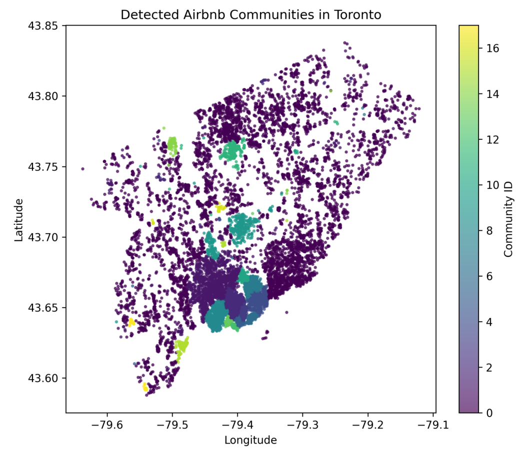
Graph	Edges	Avg. degree	Median degree	Components	Largest comp.
Graph A: Spatial only	1,552,469	196.40	66	100	87.8%
Graph B: Spatial + shared host	1,555,991	196.85	67	53	98.7%
Graph C: Full graph	1,607,040	203.31	73	1	100.0%

Table 4: Community detection results.

Graph	Algorithm	Communities	Modularity	Largest community	Median size
Graph A	Louvain	134	0.8051	1,408	9.5
Graph A	Leiden	136	0.8051	1,362	10.0
Graph B	Louvain	91	0.8069	1,429	7.0
Graph B	Leiden	87	0.8066	1,439	6.0
Graph C	Louvain	16	0.7927	6,817	686.5
Graph C	Leiden	18	0.7934	6,427	575.0

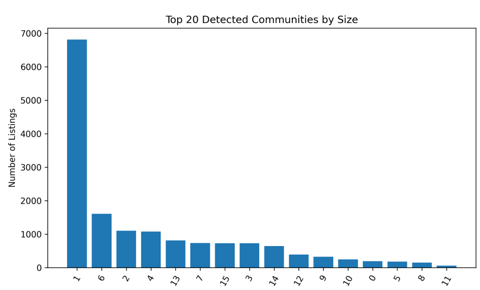


(a) Graph C Louvain communities.

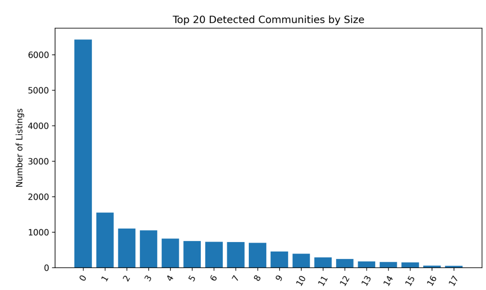


(b) Graph C Leiden communities.

Figure 3: Detected communities in the full Airbnb listing network. Community IDs are arbitrary and should not be directly matched by colour across algorithms.

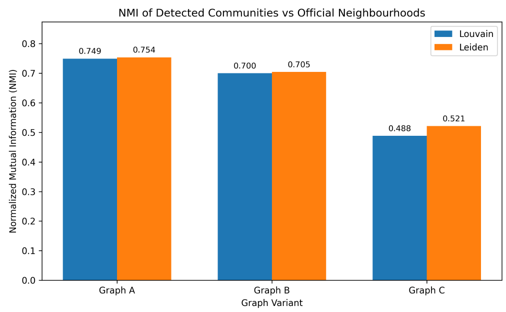


(a) Louvain community sizes.

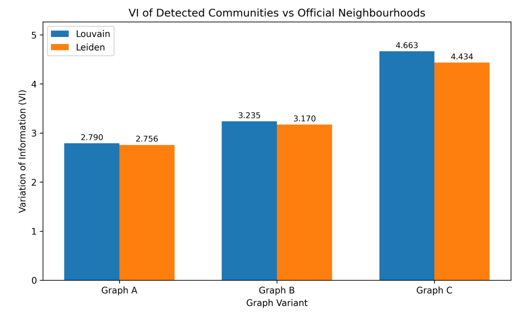


(b) Leiden community sizes.

Figure 4: Community-size distributions for Graph C. Both methods produce one dominant community and several smaller communities, but Leiden splits the network slightly more.



(a) **NMI comparison. Higher is better.**



(b) **VI comparison. Lower is better.**

Figure 5: Alignment between detected communities and official Toronto neighbourhood labels for Louvain and Leiden across the three graph variants. Leiden gives slightly stronger alignment than Louvain in all three graph variants, but both algorithms show the same overall trend.

Table 5: Community alignment with official neighbourhoods.

Graph	Algorithm	NMI	VI
A	Louvain	0.7494	2.7895
A	Leiden	0.7538	2.7562
B	Louvain	0.6999	3.2351
B	Leiden	0.7045	3.1703
C	Louvain	0.4881	4.6632
C	Leiden	0.5213	4.4341

layout, and Figure 8 places them in a single market landscape of price against geographic spread.

Three qualitatively different kinds of segment emerge:

- **A broad background market (Community 1).** The largest community contains 6,817 listings (43% of the network), spans 125 of the 140 official neighbourhoods, and has a median price of only \$95. It is dominated by entire-home listings but is spread across almost the whole city. This is the "rest of Toronto" segment that the attribute-similarity bridges merge together, rather than a single neighbourhood cluster.
- **Premium downtown clusters (Communities 3, 4, 7, 15).** These are geographically tight, high-price segments concentrated in the Waterfront/Islands and Bay Street Corridor area. Community 3, for example, contains 731 listings almost entirely inside a single official neighbourhood (Waterfront Communities–The Island) at a median price of \$194, and Community 4 holds 1,080 listings across only five neighbourhoods at a median of \$201. These behave like a downtown short-stay tourism market.
- **Budget private-room pockets (Communities 5 and 11).** These are the only private-room–dominated communities. Community 5 (York University Heights, 181 listings, 48 hosts, median \$45) and Community 11 (Bedford Park, 58 listings, median \$51) sit far below the rest of the market and resemble student- and suburb-oriented budget rentals rather than tourism listings.

Figure 9 makes the local-versus-citywide distinction explicit by plotting listings-per-neighbourhood concentration against community size. The premium downtown clusters sit high (a tight local footprint), while Community 1 sits low (a thin citywide spread). This is the key qualitative outcome of the analysis: community detection on the full graph recovers a small number of *functional* market segments — a dominant citywide background market, a handful of premium downtown clusters, and a couple of budget private-room pockets — that cut across the 140 administrative neighbourhood labels. These price differences across segments are also what give community membership its modest predictive value in the price model.

5.5 Price Prediction Results

Table 7 compares the baseline price model with the community-enhanced model. Adding Graph C Louvain community membership increases test R^2 from 0.6398 to 0.6405. It also decreases log-price MAE from 0.3204 to 0.3197 and decreases approximate dollar MAE from \$54.59 to \$54.49.

The improvement is small but measurable. This suggests that detected communities contain some price-relevant information beyond official neighbourhood labels, but most explainable variation is already captured by listing attributes and official neighbourhood effects. Therefore, community membership is best interpreted as a supplementary market-structure feature rather than a replacement for standard predictors.

5.6 Discussion

The results show a clear trade-off between geographic interpretability and market-relationship richness. Graph A is easiest to interpret geographically and aligns best with official neighbourhoods. Graph C is more expressive because it includes host and attribute information, but it no longer corresponds closely to administrative boundaries. This makes Graph C useful for discovering broader Airbnb market segments, while Graph A is better for studying neighbourhood-level spatial structure.

The Louvain-Leiden comparison also strengthens the findings. Leiden finds slightly more Graph C communities and slightly higher

Table 6: All 16 Graph C Louvain communities as Airbnb market segments.

Community	Listings	Median price	Neigh. spanned	Dominant neighbourhood	Dominant room type
1	6,817	$95	125	South Riverdale	Entire home/apt
6	1,610	$120	53	Dovercourt-Wallace Emerson-Junction	Entire home/apt
2	1,103	$110	20	Kensington-Chinatown	Entire home/apt
4	1,080	$201	5	Waterfront Communities-The Island	Entire home/apt
13	813	$161	16	Moss Park	Entire home/apt
7	738	$194	3	Waterfront Communities-The Island	Entire home/apt
15	732	$167	15	Bay Street Corridor	Entire home/apt
3	731	$194	1	Waterfront Communities-The Island	Entire home/apt
14	642	$119	30	Willowdale East	Entire home/apt
12	389	$180	6	Annex	Entire home/apt
9	328	$125	14	Annex	Entire home/apt
10	245	$192	1	Niagara	Entire home/apt
0	192	$176	3	Niagara	Entire home/apt
5	181	$45	7	York University Heights	Private room
8	150	$160	5	Mimico (Humber Bay Shores)	Entire home/apt
11	58	$51	3	Bedford Park-Nortown	Private room

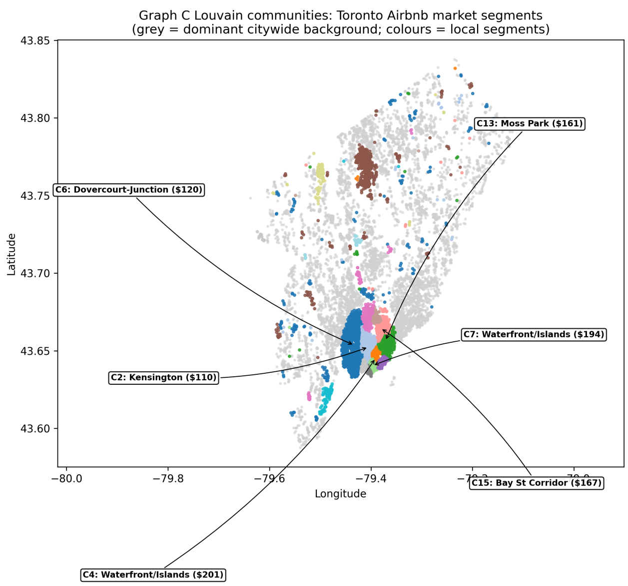


Figure 6: Spatial layout of the Graph C Louvain communities. The dominant citywide background market (Community 1) is shown in grey; the eight largest structured segments are coloured and listed in the legend with their dominant neighbourhood and median price. Premium clusters concentrate in the downtown waterfront and Bay Street area, while the grey background spreads across the whole city.

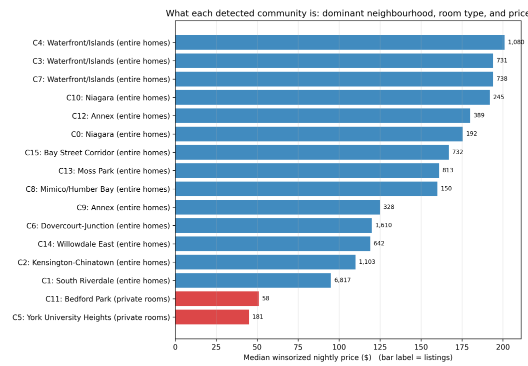


Figure 7: Each detected community as a market segment, labelled by dominant neighbourhood and room type. Bar length is median price; the number at the end of each bar is the community size. Blue bars are entire-home segments; red bars are private-room segments.

Table 7: Price prediction results on the test set.

Model	R^2	MAE	RMSE
Baseline	0.6398	$54.59	$101.36
Baseline + community	0.6405	$54.49	$101.02

modularity, but the difference is small. This suggests that the detected market structure is not an artefact of a single community detection method. However, community IDs should not be compared directly across algorithms because IDs are arbitrary labels assigned after partitioning.

A limitation is that the graph depends on design choices such as the 500 meter spatial radius, the number of nearest neighbours, and the edge-weight coefficients. Different parameter values could produce different communities. Another limitation is that price prediction uses a standard train-test split rather than spatial cross-validation. Future work could test parameter sensitivity, use temporal snapshots, or include additional text and review features.

5.7 Runtime and Reproducibility

The experiments were designed to be reproducible from the cleaned CSV file. The final code package contains the main experiment script, a helper script for Graph C Leiden figures, the generated CSV outputs, and the final figures used in the report. The main script constructs the three graph variants, computes graph statistics, runs Louvain and optional Leiden community detection, evaluates neighbourhood alignment, and trains the price models. The helper script was added only to regenerate Graph C Leiden figures without rerunning the full pipeline.

Table 8 shows that graph construction was more expensive than community detection for Graph B and Graph C. Graph B required additional shared-host nearest-neighbour searches, while Graph C

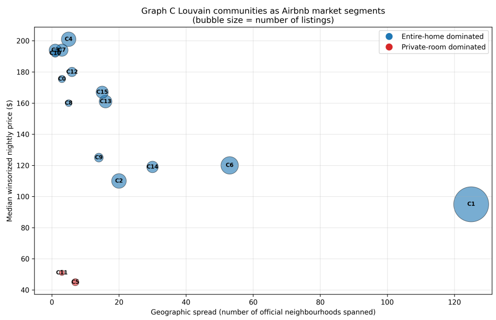


Figure 8: The detected communities in a single market landscape: geographic spread (neighbourhoods spanned) against median price, with bubble size proportional to listings. The citywide background market (Community 1) sits at the lower right; tight premium downtown clusters sit at the upper left; budget private-room segments sit at the bottom.

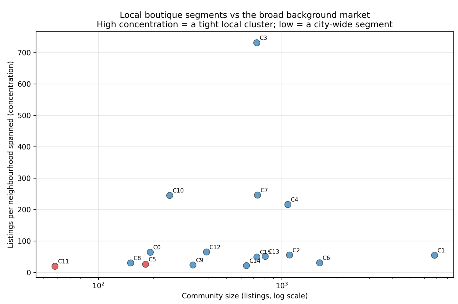


Figure 9: Local boutique segments versus the broad background market. Listings-per-neighbourhood concentration (vertical axis) separates geographically tight communities from the thin, citywide Community 1.

Table 8: Approximate runtime for graph construction and community detection.

Graph	Build (s)	Louvain (s)	Leiden (s)
A	8.68	22.11	6.83
B	214.02	33.43	7.31
C	203.29	35.44	7.88

required both shared-host and attribute-similarity edge generation. Leiden was faster than Louvain in these runs after the graph had already been constructed. However, runtime comparisons should be interpreted cautiously because the two implementations use different libraries and internal optimizations.

The most important implementation issue was avoiding dense shared-host cliques. A naive approach would connect every pair of listings owned by the same host. For a host with m listings, this creates $m(m-1)/2$ host edges, which can grow quickly for large hosts. The final implementation instead connects each listing to at most five nearest same-host listings. This keeps the graph sparse while still encoding the fact that same-host listings are related. This choice also makes the methodology easier to justify: same-host listings that are geographically close are more likely to represent coordinated local host strategy than all listings owned by the same host across the entire city.

5.8 Summary of Research Questions

The experimental results answer the three research questions as follows. For RQ1, adding non-spatial relationships clearly changes graph connectivity. Graph A has 100 connected components, Graph B has 53, and Graph C has only one. This means that host and attribute relationships act as bridges between otherwise separated spatial clusters. For RQ2, detected communities partially align with official neighbourhoods, but the level of agreement depends strongly on graph design. Spatial-only communities have high NMI, while full-network communities have lower NMI and higher VI, showing that they capture broader market structure. For RQ3, community membership improves price prediction slightly. The effect is small, but it is consistent with the community-level price differences observed in Section 5.4.

These findings are useful because they show both the strength and limitation of network features. The network representation reveals clusters that are not visible from administrative labels alone, but those clusters do not dramatically change price prediction once standard listing attributes and neighbourhood labels are already included. Therefore, the network approach is most valuable for exploratory market structure analysis and supplementary modelling, rather than as a stand-alone pricing model.

6 Conclusion

This project represented Toronto Airbnb listings as an information network in order to identify data-driven market communities rather than relying only on official neighbourhood labels. The spatial-only graph produced many neighbourhood-like communities, while the full graph became connected and produced fewer, larger communities that diverged from official neighbourhood boundaries. These full-graph communities correspond to interpretable Airbnb market segments: a dominant citywide background market, several tight premium downtown clusters, and a few budget private-room pockets. This suggests that Airbnb market structure is shaped not only by geography, but also by host ownership and listing similarity. Louvain and Leiden produced similar modularity scores, indicating stable community structure across algorithms.

The price analysis showed that adding Graph C community membership slightly improves predictive performance. The improvement is not large, but it indicates that network communities capture some market information not fully represented by official neighbourhood labels and listing attributes. Overall, the project shows that information-network methods are useful for understanding Airbnb markets, especially when the goal is to move beyond administrative neighbourhoods and identify data-driven rental market segments.

Code Availability

The project source code, cleaned dataset, experiment scripts, generated tables, figures, and final report files are available in the project GitHub repository:

https://github.com/souravC01/EECS4414-Airbnb-Network-Analysis

Acknowledgments

The authors used generative AI tools for language refinement, LaTeX formatting support, and debugging assistance. All project ideas, data processing decisions, experiments, interpretations, and final content were reviewed, revised, and approved by the authors.

References

[1] Inside Airbnb. Get the Data: Airbnb Listings Data (Toronto, November 2025 snapshot). https://insideairbnb.com/get-the-data/, 2025.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

[4] Timm Teubner. The web of host–guest connections on Airbnb: a network perspective. *Journal of Systems and Information Technology*, 20(3):262–277, 2018.

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

[6] Marina Meila. Comparing clusterings: an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[7] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[8] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, 2008.

[9] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Dan Wang and Juan L. Nicolau. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, 62:120–131, 2017.

[11] Chris Gibbs, Daniel Guttentag, Ulrike Gretzel, Jamie Morton, and Antony Goodwill. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1):46–56, 2018.