Report: Mining Bi-partite Graphs

By : Alex Szpakiewicz, Léonard Roussard, Ruben Leon, Océan Spiess

Introduction

This report presents the results obtained from analyzing bi-partite graphs using the Neo4j tool. The main goal is to study the similarity and link prediction between users and tourist locations.

2.1 Similarity

2.1.1 The two French users who left the most reviews

```
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, SUM(r.NB) AS totalReviews
ORDER BY totalReviews DESC
LIMIT 2
RETURN u.id AS userId, totalReviews
```

Results:

userId	totalReviews
70	5428
76	3420

Explanation:

1. **Query Purpose**: The Cypher query is designed to find the two French users who have left the most reviews in the database. It specifically looks at users from France and counts their total number of reviews across all areas.

2. **Top Reviewers**:

    - User ID 70: Left 5,428 reviews
    - User ID 76: Left 3,420 reviews

3. **Volume of Reviews**:

    - Both users have left an exceptionally high number of reviews, with User 70 having about 59% more reviews than User 76.
    - These numbers are significantly higher than what one might expect from typical users, suggesting these are highly active contributors to the platform.

4. **Potential Insights**:

    - **Super Users**: These two users could be classified as "super users" or "power users" of the platform. Their high level of engagement is valuable for the platform but may also warrant further investigation.

    - **Data Quality Consideration**: Such a high number of reviews from individual users might raise questions about the quality and diversity of the reviews. It's worth considering whether these users are providing in-depth, varied reviews or if there's any pattern or potential bias in their contributions.

    - **User Motivation**: It would be interesting to understand what motivates these users to leave so many reviews. Are they professional reviewers, travel bloggers, or simply very enthusiastic travelers?

    - **Geographic Coverage**: Given the high number of reviews, these users likely have visited many different areas. It could be valuable to analyze the geographic spread of their reviews to understand their travel patterns.

5. **Platform Implications**:

    - These super users might have a significant influence on the overall ratings and perceptions of various locations on the platform.
    - Their extensive contributions could be leveraged for marketing or user engagement strategies.

6. **Further Investigation**:

    - It would be beneficial to look at the distribution of reviews across different areas for these users.
    - Analyzing the content and ratings of their reviews could provide insights into their reviewing patterns and potential biases.
    - Comparing their activity to the average French user or users from other countries could offer more context.

7. **Data Integrity**:

    - While these could be legitimate super users, it's also worth verifying that these high numbers are not the result of any data anomalies or system issues.

In conclusion, these results highlight two exceptionally active French users on the platform. Their high level of engagement presents both opportunities (in terms of user insights and platform promotion) and potential concerns (regarding review diversity and data quality) that would be worth exploring further.

2.1.2 Jaccard Similarity between distinct areas

```
MATCH (u1: User{id: 70}) – [:review] –> (a: Area_4)
WITH u1, collect(distinct id(a)) AS ulAreas
MATCH (u2: Userfid: 76}) – [:review] →> (a: Area_4)
WITH u1, ulAreas, u2, collect (distinct id(a)) AS u2Areas
RETURN u1. id AS u1, u2. id AS u2,
gds. similarity. jaccard(uAreas, u2Areas) AS similarity;
```

Results:

user1Id	user2Id	jaccardSimilarity
70	76	0.08359133126934984

Explanation:

1. **Query Purpose**: This query calculates the Jaccard similarity between the areas reviewed by two specific French users (User 70 and User 76) who were identified in the previous query as the most active reviewers.

2. **Jaccard Similarity**:

    - The Jaccard similarity coefficient is a measure of the overlap between two sets.
    - It's calculated as the size of the intersection divided by the size of the union of the two sets.
    - The result ranges from 0 (no overlap) to 1 (complete overlap).

3. **Result**:

    - Jaccard Similarity: 0.08359133126934984 (approximately 0.0836 or 8.36%)

4. **Interpretation**:

    - The similarity score of about 0.0836 indicates a relatively low overlap between the areas reviewed by User 70 and User 76.
    - This means that despite both users being very active reviewers, they have reviewed mostly different areas.

5. **Implications**: a) **Diverse Coverage**:

    - These top reviewers are contributing to a wide range of different areas, which is beneficial for the platform's overall coverage.

    - They are not concentrating their reviews on the same locations.

    b) **Different Travel Patterns or Preferences**:

    - User 70 and User 76 likely have different travel habits, interests, or geographic focuses.
    - This could indicate diverse perspectives among even the most active users.

    c) **Platform Diversity**:

    - The low similarity suggests that the platform benefits from diverse inputs even from its most active users.
    - This diversity can provide a richer set of reviews across different locations.

    d) **Complementary Information**:

    - The reviews from these users are likely complementary rather than redundant, adding value to the platform's content.

6. **Further Considerations**:

    - It would be interesting to investigate if this low similarity is common among other pairs of users or if it's unique to these top reviewers.
    - Analyzing the specific areas each user reviews could provide insights into their travel preferences or specializations.
    - Comparing this similarity score with the average similarity between random pairs of users could offer context on how unique or typical this pattern is.

7. **Potential Follow-up Analyses**:

    - Examine the types of areas (e.g., cities, rural areas, tourist hotspots) each user tends to review.
    - Investigate if there are any common characteristics in the small overlap of areas they both reviewed.
    - Look into temporal patterns - do they review different areas in different time periods?

In conclusion, while User 70 and User 76 are both highly active reviewers, they appear to be contributing to the platform in quite different ways, reviewing largely distinct sets of areas. This diversity in their reviewing patterns adds richness to the platform's content and suggests that even among top contributors, there's a wide range of travel experiences being shared.

2.1.3 Similarity for the two French users who visited the most distinct areas

```
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, COUNT(DISTINCT a) AS distinctAreas
ORDER BY distinctAreas DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[:review]->(a1:Area_4)
MATCH (u2)-[:review]->(a2:Area_4)
WITH u1, u2, COLLECT(DISTINCT a1) AS areas1, COLLECT(DISTINCT a2) AS areas2
WITH u1, u2, areas1, areas2,
     [x IN areas1 WHERE x IN areas2] AS intersection,
     areas1 + [x IN areas2 WHERE NOT x IN areas1] AS union
RETURN u1.id AS user1Id, u2.id AS user2Id,
       SIZE(intersection)*1.0 / SIZE(union) AS jaccardSimilarity
```

Results:

user1Id	user2Id	jaccardSimilarity
2639	387312	0.04791666666666667

Explanation:

1. **Query Purpose**: This query identifies the two French users who have visited the most distinct areas and then calculates the Jaccard similarity between the areas they've reviewed.

2. **Users Identified**:

   - User ID: 2639
   - User ID: 387312 These users have visited the highest number of distinct areas among French users in the database.

3. **Jaccard Similarity**:

   - Jaccard Similarity: 0.04791666666666667 (approximately 0.0479 or 4.79%)

4. **Interpretation**: a) **Very Low Similarity**:

   - The Jaccard similarity of about 4.79% indicates an extremely low overlap between the areas visited by these two users.
   - This means that despite both users being among the most widely traveled (in terms of distinct areas visited), they have mostly been to different places.

   b) **Diverse Exploration Patterns**:

   - These users, while both extensive travelers, have very different travel patterns or preferences.
   - They contribute to the platform's diversity by providing reviews for largely different sets of areas.

   c) **Breadth of Coverage**:

   - The low similarity suggests that these users collectively provide a very broad coverage of different areas on the platform.
   - This is beneficial for the platform as it offers diverse perspectives and information on a wide range of locations.

   d) **Potential for Complementary Insights**:

   - With such different travel histories, these users likely offer complementary insights and experiences, enriching the overall content of the platform.

5. **Comparison with Previous Results**:

   - The similarity here (4.79%) is even lower than the similarity found between the two most active reviewers in the previous query (8.36%).
   - This suggests that users who visit many distinct areas tend to have even more diverse travel patterns compared to those who simply review frequently.

6. **Implications for the Platform**:

   - The platform benefits from having users with such diverse travel experiences.
   - It indicates that even among the most well-traveled users, there's a significant variety in the places they visit and review.

7. **Further Considerations**:

   - It would be interesting to investigate the types of areas each user tends to visit. Are they focusing on different regions, types of destinations (urban vs. rural), or perhaps different types of attractions?
   - Analyzing the small overlap in their visited areas could provide insights into popular or notable locations that attract even diverse travelers.
   - Examining the total number of distinct areas visited by each user could give context to their travel breadth.

8. **Potential Follow-up Analyses**:

   - Compare these users' travel patterns with those of average users to understand how exceptional their diversity is.
   - Investigate if there are any common characteristics (e.g., seasonality, type of location) in the areas they both visited.
   - Analyze the content of their reviews to see if they focus on different aspects of the areas they visit.

In conclusion, these results highlight two French users with exceptionally diverse travel patterns. Their low similarity score indicates that they contribute unique and varied content to the platform, covering a wide range of distinct areas. This diversity is valuable for providing comprehensive coverage and varied perspectives on different locations, enhancing the overall quality and breadth of information available on the platform.

2.1.4 Explanation of differences

1. **Comparison of Results**:
    - Most active reviewers (User 70 and 76): Jaccard similarity of 0.08359133126934984 (≈8.36%)
    - Users who visited most distinct areas (User 2639 and 387312): Jaccard similarity of 0.04791666666666667 (≈4.79%)

2. **Key Difference**: The similarity between the users who visited the most distinct areas is notably lower (by about 3.57 percentage points) than the similarity between the most active reviewers.

3. **Interpretation of the Difference**:

    a) **Review Quantity vs. Area Diversity**:

    - The most active reviewers (70 and 76) may have a higher overlap because they're reviewing more frequently, potentially including some common popular areas.
    - Users who visited the most distinct areas (2639 and 387312) seem to have more diverse travel patterns, resulting in less overlap.

    b) **Depth vs. Breadth**:

    - Active reviewers might be providing multiple reviews for the same areas, leading to higher review counts but potentially less geographic diversity.
    - Users visiting many distinct areas prioritize breadth of travel, resulting in less overlap in their experiences.

    c) **Travel Styles**:

    - The most active reviewers might focus on thoroughly exploring and repeatedly reviewing certain regions or types of destinations.
    - Users visiting many distinct areas likely have a travel style that emphasizes exploring new and different locations each time.

    d) **Platform Usage Patterns**:

    - Frequent reviewers might be more likely to review popular or easily accessible locations, increasing the chance of overlap.
    - Users visiting many areas might be more inclined to seek out and review unique or less common destinations.

4. **Implications**:
    - The platform benefits from both types of users:
        - Active reviewers provide depth and potentially more detailed insights into specific areas.
        - Users visiting many distinct areas contribute to the platform's geographic breadth and diversity.
    - The lower similarity among users visiting many areas suggests that the platform has good coverage of diverse locations, not just concentrated on popular spots.

5. **Value to the Platform**:
    - The combination of these user types enhances the overall quality of the platform:
        - Depth from frequent reviewers
        - Breadth from diverse travelers
    - This diversity in user behavior contributes to a more comprehensive and varied set of reviews and experiences.

6. **Considerations for Analysis**:
    - The difference in similarity scores highlights the importance of considering different metrics (review frequency vs. area diversity) when analyzing user behavior and contributions.
    - It suggests that the platform successfully attracts and retains users with varying travel patterns and reviewing habits.

In conclusion, the difference in similarity scores between these two pairs of users reflects distinct user behaviors and contributions to the platform. The lower similarity among users who visit many distinct areas underscores the platform's ability to capture diverse travel experiences. Meanwhile, the slightly higher similarity among frequent reviewers suggests a balance between focused, in-depth coverage and broad, diverse experiences. This combination enriches the platform's content and appeal to a wide range of users and travelers.

2.1.5 Overlap and comparison with Jaccard

```
// For top reviewers by total reviews
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, SUM(r.NB) AS totalReviews
ORDER BY totalReviews DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[:review]->(a1:Area_4)
MATCH (u2)-[:review]->(a2:Area_4)
WITH u1, u2, COLLECT(DISTINCT a1) AS areas1, COLLECT(DISTINCT a2) AS areas2
WITH u1, u2, areas1, areas2,
     [x IN areas1 WHERE x IN areas2] AS intersection,
     SIZE(areas1) AS size1, SIZE(areas2) AS size2
RETURN u1.id AS user1Id, u2.id AS user2Id,
       SIZE(intersection)*1.0 / CASE WHEN size1 < size2 THEN size1 ELSE size2 END AS
overlapSimilarity

UNION

// For top reviewers by distinct areas
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, COUNT(DISTINCT a) AS distinctAreas
ORDER BY distinctAreas DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[:review]->(a1:Area_4)
MATCH (u2)-[:review]->(a2:Area_4)
WITH u1, u2, COLLECT(DISTINCT a1) AS areas1, COLLECT(DISTINCT a2) AS areas2
WITH u1, u2, areas1, areas2,
     [x IN areas1 WHERE x IN areas2] AS intersection,
     SIZE(areas1) AS size1, SIZE(areas2) AS size2
RETURN u1.id AS user1Id, u2.id AS user2Id,
       SIZE(intersection)*1.0 / CASE WHEN size1 < size2 THEN size1 ELSE size2 END AS
overlapSimilarity
```

Results:

user1Id	user2Id	overlapSimilarity
70	76	0.1656441717791411
2639	387312	0.0931174089068826

Explanation:

1. **Query Purpose**: This query calculates the overlap similarity for two sets of user pairs: a) The two French users who left the most reviews (User 70 and User 76) b) The two French users who visited the most distinct areas (User 2639 and User 387312)

2. **Overlap Similarity Measure**:

   - The overlap similarity is calculated as the size of the intersection divided by the size of the smaller set.
   - This measure focuses on how much the smaller set is contained within the larger set.

3. **Results**: a) Top reviewers by total reviews (User 70 and User 76):

    - Overlap Similarity: 0.1656441717791411 (≈16.56%)

    b) Top reviewers by distinct areas (User 2639 and User 387312):

    - Overlap Similarity: 0.0931174089068826 (≈9.31%)

4. **Interpretation**:

    a) **Comparison to Jaccard Similarity**:

    - For Users 70 and 76, the overlap similarity (16.56%) is higher than their Jaccard similarity (8.36% from previous results).
    - For Users 2639 and 387312, the overlap similarity (9.31%) is also higher than their Jaccard similarity (4.79% from previous results).
    - This increase is expected because overlap similarity focuses on the smaller set, while Jaccard considers the union of both sets.

    b) **Top Reviewers by Total Reviews**:

    - The 16.56% overlap suggests that the user with fewer reviewed areas has about 1/6 of their reviewed areas in common with the other user.
    - This indicates a moderate level of commonality in their reviewed locations, despite their high review counts.

    c) **Top Reviewers by Distinct Areas**:

    - The 9.31% overlap shows that the user who visited fewer distinct areas shares about 1/10 of their visited areas with the other user.
    - This lower overlap reinforces that these users, while both extensive travelers, have quite different travel patterns.

    d) **Comparison Between User Pairs**:

    - The higher overlap for the top reviewers by total reviews (16.56% vs 9.31%) suggests that frequent reviewers are more likely to have common areas in their reviews compared to those who visit many distinct areas.
    - This could indicate that frequent reviewers might focus more on popular or easily accessible locations, leading to more overlap.

5. **Implications**:

    a) **Review Patterns**:

    - Frequent reviewers (70 and 76) show more commonality in their reviewed areas, possibly due to focusing on popular destinations or having similar travel preferences.
    - Users visiting many distinct areas (2639 and 387312) demonstrate more diverse and unique travel patterns, with less overlap in their experiences.

    b) **Platform Diversity**:

    - The platform benefits from both types of users:
        - Frequent reviewers provide depth and potentially multiple perspectives on common areas.
        - Users visiting many distinct areas contribute to the breadth of coverage across different locations.

    c) **User Behavior Insights**:

    - The difference in overlap between the two pairs suggests varying approaches to travel and reviewing:
        - Some users might prefer revisiting and extensively reviewing certain areas.
        - Others might prioritize exploring and reviewing new, diverse locations.

6. **Further Considerations**:

    - It would be interesting to investigate the nature of the overlapping areas for each pair. Are they popular tourist destinations, or do they represent more niche locations?
    - Analyzing the non-overlapping areas could provide insights into the unique contributions of each user to the platform's content.

In conclusion, these results highlight the different patterns of user engagement on the platform. The higher overlap among frequent reviewers suggests a tendency to cover some common ground, possibly popular or significant locations. In contrast, the lower overlap among users visiting many distinct areas underscores the value of having diverse contributors who collectively provide a wide-ranging view of different travel destinations. This diversity in user behavior enriches the platform's content, offering both in-depth coverage of certain areas and broad exploration of varied locations.

2.1.6 Euclidean and Cosine similarities based on the number of reviews (NB)

```
// For top reviewers by total reviews
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, SUM(r.NB) AS totalReviews
ORDER BY totalReviews DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[r1:review]->(a:Area_4)
MATCH (u2)-[r2:review]->(a:Area_4)
WITH u1, u2,
     COLLECT({areaId: a.gid, NB: toFloat(r1.NB)}) AS reviews1,
     COLLECT({areaId: a.gid, NB: toFloat(r2.NB)}) AS reviews2
WITH u1, u2, reviews1, reviews2,
     SQRT(REDUCE(s = 0.0, r IN reviews1 |
       s + (r.NB - CASE WHEN r.areaId IN [rev IN reviews2 | rev.areaId]
                        THEN [rev IN reviews2 WHERE rev.areaId = r.areaId | rev.NB][0]
                        ELSE 0.0 END)^2
     )) AS euclideanDistance,
     REDUCE(dotProduct = 0.0, r IN reviews1 |
       dotProduct + r.NB * CASE WHEN r.areaId IN [rev IN reviews2 | rev.areaId]
                                THEN [rev IN reviews2 WHERE rev.areaId = r.areaId | rev.NB]
[0]
                                ELSE 0.0 END
     ) / (SQRT(REDUCE(s = 0.0, r IN reviews1 | s + r.NB^2)) *
          SQRT(REDUCE(s = 0.0, r IN reviews2 | s + r.NB^2))) AS cosineSimilarity
RETURN u1.id AS user1Id, u2.id AS user2Id,
       euclideanDistance AS euclideanDistance,
       cosineSimilarity AS cosineSimilarity

UNION

// For top reviewers by distinct areas
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, COUNT(DISTINCT a) AS distinctAreas
ORDER BY distinctAreas DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[r1:review]->(a:Area_4)
MATCH (u2)-[r2:review]->(a:Area_4)
WITH u1, u2,
     COLLECT({areaId: a.gid, NB: toFloat(r1.NB)}) AS reviews1,
     COLLECT({areaId: a.gid, NB: toFloat(r2.NB)}) AS reviews2
WITH u1, u2, reviews1, reviews2,
     SQRT(REDUCE(s = 0.0, r IN reviews1 |
       s + (r.NB - CASE WHEN r.areaId IN [rev IN reviews2 | rev.areaId]
                        THEN [rev IN reviews2 WHERE rev.areaId = r.areaId | rev.NB][0]
                        ELSE 0.0 END)^2
     )) AS euclideanDistance,
     REDUCE(dotProduct = 0.0, r IN reviews1 |
       dotProduct + r.NB * CASE WHEN r.areaId IN [rev IN reviews2 | rev.areaId]
```

```
                                THEN [rev IN reviews2 WHERE rev.areaId = r.areaId | rev.NB]
[0]
                                ELSE 0.0 END
     ) / (SQRT(REDUCE(s = 0.0, r IN reviews1 | s + r.NB^2)) *
          SQRT(REDUCE(s = 0.0, r IN reviews2 | s + r.NB^2))) AS cosineSimilarity
RETURN u1.id AS user1Id, u2.id AS user2Id,
       euclideanDistance AS euclideanDistance,
       cosineSimilarity AS cosineSimilarity
```

Results:

user1Id	user2Id	euclideanDistance	cosineSimilarity
70	76	702.6478492103993	0.5498294794404108
2639	387312	29.46183972531247	0. 7926268998711051

Explanation:

1. **Query Purpose**: This query calculates two similarity measures (Euclidean distance and Cosine similarity) based on the number of reviews (NB) for two sets of user pairs: a) The two French users who left the most reviews (User 70 and User 76) b) The two French users who visited the most distinct areas (User 2639 and User 387312)

2. **Similarity Measures**:

   - Euclidean Distance: Measures the straight-line distance between two points in multi-dimensional space. Lower values indicate more similarity.
   - Cosine Similarity: Measures the cosine of the angle between two vectors. Values range from -1 to 1, with 1 indicating perfect similarity.

3. **Results**: a) Top reviewers by total reviews (User 70 and User 76):

   - Euclidean Distance: 702.6478492103993
   - Cosine Similarity: 0.5498294794404108

   b) Top reviewers by distinct areas (User 2639 and User 387312):

   - Euclidean Distance: 29.46183972531247
   - Cosine Similarity: 0.7926268998711051

4. **Interpretation**:

   a) **Euclidean Distance**:

   - The distance is much larger for the top reviewers by total reviews (702.65) compared to the top reviewers by distinct areas (29.46).
   - This suggests that Users 70 and 76 have more divergent review patterns in terms of the number of reviews they leave for each area.
   - Users 2639 and 387312 have a much smaller Euclidean distance, indicating more similar numbers of reviews across the areas they've both visited.

   b) **Cosine Similarity**:

   - Both pairs show positive cosine similarity, indicating some level of similarity in their review patterns.
   - The top reviewers by distinct areas (0.7926) have a higher cosine similarity than the top reviewers by total reviews (0.5498).
   - This suggests that Users 2639 and 387312 have more similar proportions of reviews across the areas they've both visited, even if the absolute numbers differ.

   c) **Comparison Between User Pairs**:

   - The top reviewers by total reviews (70 and 76) show less similarity in both measures compared to the top reviewers by distinct areas (2639 and 387312).

- This could indicate that while Users 70 and 76 review more overall, they have more divergent focuses or preferences in terms of which areas they review more frequently.
- Users 2639 and 387312, despite visiting many distinct areas, seem to have more similar patterns in how they distribute their reviews across the areas they've both visited.

5. **Implications**:

a) **Review Behavior**:

- Frequent reviewers (70 and 76) might have more individualized preferences or focuses, leading to larger differences in their review counts for specific areas.
- Users visiting many distinct areas (2639 and 387312) seem to have more consistent reviewing behavior across the areas they both visit.

b) **User Profiling**:

- These similarity measures could be useful for user profiling and recommendation systems on the platform.
- The higher similarity between Users 2639 and 387312 suggests they might have more similar travel preferences or reviewing styles.

c) **Content Distribution**:

- The platform benefits from the diversity of the top reviewers (70 and 76), as they likely provide varied depth of coverage across different areas.
- The more consistent behavior of Users 2639 and 387312 might provide a more balanced coverage across a wide range of areas.

6. **Further Considerations**:

- It would be interesting to investigate why the top reviewers by total reviews have such a large Euclidean distance. Are there specific areas where their review counts differ dramatically?
- For the users with high distinct area counts, analyzing the areas they both visited versus those unique to each could provide insights into travel patterns and preferences.

In conclusion, these results reveal different patterns of user engagement and review behavior. The top reviewers by total reviews show more divergent patterns in their review counts, suggesting individualized focus areas or preferences. In contrast, the users who visited the most distinct areas demonstrate more similar reviewing patterns across shared locations, despite their wide-ranging travels. This diversity in user behavior contributes to the richness of the platform's content, offering both varied depth in specific areas and consistent breadth across many locations.

2.1.7 Similarities based on ratings

```
// Find top French reviewers by total reviews and distinct areas
MATCH (u:User {country: 'France'})
WITH u,
     SIZE([(u)-[:review]->() | 1]) AS totalReviews,
     SIZE([(u)-[:review]->(:Area_4) | 1]) AS distinctAreas
WITH u, totalReviews, distinctAreas
ORDER BY totalReviews DESC, distinctAreas DESC
LIMIT 4
WITH COLLECT({user: u, total: totalReviews, distinct: distinctAreas}) AS topReviewers

// Process both pairs
WITH topReviewers[0].user AS topTotalUser1,
     topReviewers[1].user AS topTotalUser2,
     topReviewers[2].user AS topDistinctUser1,
     topReviewers[3].user AS topDistinctUser2

// Calculate similarities for total reviews pair
MATCH (topTotalUser1)-[r1:review]->(a:Area_4)<-[r2:review]-(topTotalUser2)
WITH topTotalUser1, topTotalUser2, topDistinctUser1, topDistinctUser2,
     COLLECT({rating1: toFloat(r1.rating), rating2: toFloat(r2.rating)}) AS
```

```
totalCommonReviews

// Calculate similarities for distinct areas pair
MATCH (topDistinctUser1)-[r1:review]->(a:Area_4)<-[r2:review]-(topDistinctUser2)
WITH topTotalUser1, topTotalUser2, topDistinctUser1, topDistinctUser2,
     totalCommonReviews,
     COLLECT({rating1: toFloat(r1.rating), rating2: toFloat(r2.rating)}) AS
distinctCommonReviews

// Calculate Euclidean distance and Cosine similarity for both pairs
WITH ['Total Reviews', 'Distinct Areas'] AS pairTypes,
     [totalCommonReviews, distinctCommonReviews] AS commonReviewsList,
     [topTotalUser1, topDistinctUser1] AS user1List,
     [topTotalUser2, topDistinctUser2] AS user2List
UNWIND RANGE(0, 1) AS i
WITH pairTypes[i] AS pairType,
     commonReviewsList[i] AS commonReviews,
     user1List[i] AS u1,
     user2List[i] AS u2
WITH pairType, u1, u2, commonReviews,
     SQRT(REDUCE(s = 0.0, r IN commonReviews | s + (r.rating1 - r.rating2)^2)) AS
euclideanDistance,
     CASE SIZE(commonReviews)
       WHEN 0 THEN null
       ELSE REDUCE(dot = 0.0, r IN commonReviews | dot + r.rating1 * r.rating2) /
            (SQRT(REDUCE(s = 0.0, r IN commonReviews | s + r.rating1^2)) *
             SQRT(REDUCE(s = 0.0, r IN commonReviews | s + r.rating2^2)))
     END AS cosineSimilarity

// Return results
RETURN
  pairType AS userPair,
  u1.id AS user1Id,
  u2.id AS user2Id,
  euclideanDistance,
  cosineSimilarity,
  SIZE(commonReviews) AS commonReviewsCount
```

Results:

userPair	user1Id	user2Id	euclideanDistance	cosineSimilarity	commonReviewsCount
"Distinct Areas"	2639	387312	9.342303896293679	0.9802174560172844	74
"Total Reviews"	70	76	11.636082386970404	0.9756447834632881	172

Explanation:

1. **User Pairs**:
   - "Distinct Areas" pair: Users 2639 and 387312
   - "Total Reviews" pair: Users 70 and 76
2. **Euclidean Distance**:
   - "Distinct Areas" pair: 9.342303896293679
   - "Total Reviews" pair: 11.636082386970404

   Interpretation:

   - The Euclidean distance is smaller for the "Distinct Areas" pair, indicating that these users have more similar rating patterns for the areas they've both reviewed.
   - The "Total Reviews" pair shows a slightly larger Euclidean distance, suggesting more variation in their ratings for common areas.

3. **Cosine Similarity**:

    - "Distinct Areas" pair: 0.9802174560172844
    - "Total Reviews" pair: 0.9756447834632881

    Interpretation:

    - Both pairs show very high cosine similarity (close to 1), indicating that their rating patterns are very similar in direction.
    - The "Distinct Areas" pair has a marginally higher cosine similarity, suggesting slightly more aligned rating behaviors.

4. **Common Reviews Count**:

    - "Distinct Areas" pair: 74 common reviews
    - "Total Reviews" pair: 172 common reviews

    Interpretation:

    - As expected, the "Total Reviews" pair has more than twice as many common reviews, aligning with their status as the most frequent reviewers.
    - The "Distinct Areas" pair, despite focusing on unique locations, still has a significant number of common reviews.

5. **Overall Interpretation**:

    a) High Similarity in Rating Behavior:

    - Both pairs show very high cosine similarities, indicating that users tend to rate areas in similar ways, regardless of their reviewing patterns.
    - This suggests a level of consistency in how users perceive and rate locations on the platform.

    b) Difference Between Pairs:

    - The "Distinct Areas" pair shows slightly more similar rating patterns (higher cosine similarity, lower Euclidean distance) despite having fewer common reviews.
    - This could indicate that users who explore more diverse areas might develop more consistent rating criteria.

    c) Frequency vs. Diversity:

    - The "Total Reviews" pair, despite having more common reviews, shows slightly more divergence in their ratings.
    - This could suggest that frequent reviewers might develop more nuanced or varied rating habits over time.

6. **Implications**:

    a) User Behavior:

    - Users tend to have consistent rating patterns, regardless of whether they focus on reviewing many times or exploring diverse areas.
    - This consistency could be valuable for recommendation systems and understanding user preferences.

    b) Platform Insights:

    - The high similarities suggest that the platform's rating system is being used consistently across different user types.
    - It may indicate that the platform provides clear rating criteria or that users develop similar standards for evaluating areas.

    c) Data Quality:

    - The consistency in ratings across different user types suggests good data quality and reliable user input.

7. **Considerations for Further Analysis**:

    a) Rating Distribution:

    - It would be interesting to examine the distribution of ratings for each user to see if there are any biases (e.g., tendency to give high or low ratings).

    b) Area Types:

  - Analyzing whether the similarity patterns hold across different types of areas (e.g., tourist attractions, restaurants, hotels) could provide deeper insights.

c) Temporal Analysis:

  - Investigating how these similarities change over time could reveal evolving user behaviors or preferences.

In conclusion, these results reveal a high degree of consistency in how users rate areas, regardless of their reviewing patterns. This suggests that the platform has successfully established a reliable rating system that is used similarly across different user types. The slight differences between the pairs hint at nuanced variations in rating behavior between frequent reviewers and those who explore more diverse areas, providing valuable insights for understanding user engagement and preferences on the platform.

2.1.8 Similarities for common areas only

```
// Function to calculate Euclidean distance
WITH function() {
  RETURN sqrt(sum((a.NB - b.NB)^2 for a in commonReviews1
               for b in commonReviews2
               where a.areaId = b.areaId))
} AS euclideanDistanceNB,
// Function to calculate Cosine similarity for NB
function() {
  RETURN sum(a.NB * b.NB for a in commonReviews1
             for b in commonReviews2
             where a.areaId = b.areaId) /
         (sqrt(sum(a.NB^2 for a in commonReviews1)) *
          sqrt(sum(b.NB^2 for b in commonReviews2)))
} AS cosineSimilarityNB,
// Function to calculate Euclidean distance for ratings
function() {
  RETURN sqrt(sum((a.rating - b.rating)^2 for a in commonReviews1
               for b in commonReviews2
               where a.areaId = b.areaId))
} AS euclideanDistanceRating,
// Function to calculate Cosine similarity for ratings
function() {
  RETURN sum(a.rating * b.rating for a in commonReviews1
             for b in commonReviews2
             where a.areaId = b.areaId) /
         (sqrt(sum(a.rating^2 for a in commonReviews1)) *
          sqrt(sum(b.rating^2 for b in commonReviews2)))
} AS cosineSimilarityRating

// For top reviewers by total reviews
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, SUM(r.NB) AS totalReviews
ORDER BY totalReviews DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[r1:review]->(a:Area_4)<-[r2:review]-(u2)
WITH u1, u2,
     COLLECT({areaId: a.gid, NB: r1.NB, rating: r1.rating}) AS commonReviews1,
     COLLECT({areaId: a.gid, NB: r2.NB, rating: r2.rating}) AS commonReviews2,
     euclideanDistanceNB, cosineSimilarityNB,
     euclideanDistanceRating, cosineSimilarityRating
RETURN u1.id AS user1Id, u2.id AS user2Id,
       euclideanDistanceNB() AS euclideanDistanceNB,
       cosineSimilarityNB() AS cosineSimilarityNB,
       euclideanDistanceRating() AS euclideanDistanceRating,
       cosineSimilarityRating() AS cosineSimilarityRating
```

```
UNION

// For top reviewers by distinct areas
MATCH (u:User {country: 'France'})-[r:review]->(a:Area_4)
WITH u, COUNT(DISTINCT a) AS distinctAreas
ORDER BY distinctAreas DESC
LIMIT 2
WITH COLLECT(u) AS topUsers

MATCH (u1:User) WHERE u1 IN topUsers
MATCH (u2:User) WHERE u2 IN topUsers AND u1.id < u2.id
MATCH (u1)-[r1:review]->(a:Area_4)<-[r2:review]-(u2)
WITH u1, u2,
     COLLECT({areaId: a.gid, NB: r1.NB, rating: r1.rating}) AS commonReviews1,
     COLLECT({areaId: a.gid, NB: r2.NB, rating: r2.rating}) AS commonReviews2,
     euclideanDistanceNB, cosineSimilarityNB,
     euclideanDistanceRating, cosineSimilarityRating
RETURN u1.id AS user1Id, u2.id AS user2Id,
       euclideanDistanceNB() AS euclideanDistanceNB,
       cosineSimilarityNB() AS cosineSimilarityNB,
       euclideanDistanceRating() AS euclideanDistanceRating,
       cosineSimilarityRating() AS cosineSimilarityRating
```

Results:

user1Id	user2Id	euclideanDistanceNB	cosineSimilarityNB	euclideanDistanceRating	cosineSimilarityRating
70	76	0.0016852273044407502	0.3450603868552847	0.07913845204358133	0.9756447834632883
2639	387312	0.036776060886741006	0.7320607247978977	0.09669025490136343	0.9802174560172844

Explanation:

1. Top reviewers by total reviews (User IDs 70 and 76):

   a. Euclidean Distance NB (0.0016852273044407502):

   - This value is very close to 0, indicating that these users have very similar numbers of reviews for common areas.
   - They likely have reviewed the same areas with similar frequency.

   b. Cosine Similarity NB (0.3450603868552847):

   - This value is moderate, suggesting some similarity in review patterns, but not extremely high.
   - While they may review similar areas, the proportions of reviews per area might differ.

   c. Euclidean Distance Rating (0.07913845204358133):

   - This low value indicates that their ratings for common areas are quite similar.
   - They tend to agree on the quality of the areas they've both reviewed.

   d. Cosine Similarity Rating (0.9756447834632883):

   - This high value suggests that their rating patterns are very similar.
   - They likely have very similar preferences and opinions about the areas they've reviewed.

2. Top reviewers by distinct areas (User IDs 2639 and 387312):

   a. Euclidean Distance NB (0.036776060886741006):

   - This value is higher than for the first pair, indicating more variation in the number of reviews for common areas.
   - These users might have different levels of activity in the areas they've both reviewed.

   b. Cosine Similarity NB (0.7320607247978977):

    - This higher value suggests that despite differences in review counts, the overall pattern of which areas they review more or less is quite similar.

  c. Euclidean Distance Rating (0.09669025490136343):

    - This value is slightly higher than for the first pair, but still relatively low.
    - Their ratings for common areas are quite similar, though with slightly more variation than the first pair.

  d. Cosine Similarity Rating (0.9802174560172844):

    - This very high value indicates that their rating patterns are extremely similar.
    - They likely have very similar preferences and opinions about the areas they've reviewed.

Interpretation:

1. The top reviewers by total reviews (70 and 76) have very similar numbers of reviews for common areas, but their overall review patterns (which areas they review more or less) are only moderately similar. However, their ratings are very similar, suggesting they have similar opinions about the places they've both visited.

2. The top reviewers by distinct areas (2639 and 387312) show more variation in the number of reviews they leave for common areas, but their overall patterns of which areas they review more or less are quite similar. Their ratings are even more similar than the first pair, indicating very aligned preferences and opinions.

3. Both pairs show very high similarity in their rating patterns, suggesting that frequent French reviewers (whether by total reviews or distinct areas) tend to have similar opinions about the areas they review.

4. The pair with more distinct areas reviewed (2639 and 387312) shows higher similarity in review patterns (Cosine Similarity NB) than the pair with more total reviews. This could indicate that users who explore more diverse areas tend to have more similar overall review behaviors.

5. The very high rating similarities for both pairs suggest that these top French reviewers have consistent rating behaviors and likely similar standards for evaluating areas.

This analysis provides insights into the behavior of top French reviewers, showing that while their review frequencies may vary, their opinions and rating patterns tend to be very similar, especially for users who review a wide variety of distinct areas.

2.1.9 Average similarities for Spanish users

```
MATCH (u1:User {country: 'Spain'})-[r1:review]->(a:Area_4)
WHERE r1.NB >= 5
WITH u1, COLLECT(DISTINCT a) AS areas1
MATCH (u2:User {country: 'Spain'})-[r2:review]->(a:Area_4)
WHERE r2.NB >= 5 AND u1.id < u2.id
WITH u1, u2, areas1, COLLECT(DISTINCT a) AS areas2
WHERE SIZE(areas1) >= 5 AND SIZE(areas2) >= 5
WITH u1, u2, areas1, areas2,
     [x IN areas1 WHERE x IN areas2] AS intersection
WITH u1, u2,
     SIZE(intersection)*1.0 / SIZE(areas1 + [x IN areas2 WHERE NOT x IN areas1]) AS jaccard,
     SIZE(intersection)*1.0 / apoc.coll.min([SIZE(areas1), SIZE(areas2)]) AS overlap
RETURN AVG(jaccard) AS avgJaccard, AVG(overlap) AS avgOverlap
```

Results:

Measure	Value
Average Jaccard	0.0276
Average Overlap	0.0667

Explanation: Analysis of Results for Spanish Users

1. **Similarity Measures**:
   - The Average Jaccard similarity for Spanish users is 0.0276.
   - The Average Overlap similarity for Spanish users is 0.0667.
2. **Interpretation of Jaccard Similarity**:
   - The Jaccard similarity of 0.0276 is relatively low. This measure considers the intersection of areas visited divided by the union of areas visited for pairs of users.
   - A low Jaccard score suggests that Spanish users in the dataset have quite diverse travel patterns. There is little overlap in the areas they visit when considering the total set of areas visited by pairs of users.
   - This could indicate that Spanish travelers in this sample tend to explore a wide variety of destinations rather than concentrating on a small set of common areas.
3. **Interpretation of Overlap Similarity**:
   - The Overlap similarity of 0.0667 is higher than the Jaccard similarity but still relatively low. This measure considers the intersection of areas visited divided by the size of the smaller set of areas between two users.
   - The higher Overlap score compared to Jaccard indicates that when Spanish users do visit common areas, there's a moderate level of similarity, especially when considering the smaller set of areas between pairs of users.
   - This suggests that while Spanish users have diverse overall travel patterns, there is still some commonality in destinations when focusing on the more frequently visited areas.
4. **Relationship Between Measures**:
   - The Overlap similarity being higher than the Jaccard similarity is expected and common. It indicates that users have more in common when we focus on the smaller set of destinations between pairs.
   - The difference between these measures suggests that while Spanish users might have some common preferred destinations, they also tend to explore many unique locations.
5. **Possible Explanations**:
   - Diverse travel preferences: Spanish users might have varied interests in travel destinations.
   - Exploration tendency: There could be a cultural inclination among Spanish travelers to explore a wide range of destinations rather than concentrating on popular tourist spots.
   - Geographic factors: Spain's location might influence travel patterns, possibly leading to a diverse set of destinations across Europe and beyond.
   - Economic or social factors: Differences in vacation patterns, travel budgets, or social trends could contribute to this diversity in travel choices.
6. **Implications**:
   - For the travel industry: This data suggests that Spanish travelers might be interested in a wide variety of destinations, potentially making them a diverse target market for various types of travel experiences.
   - For recommendation systems: The low Jaccard similarity but slightly higher Overlap similarity suggests that recommendation algorithms for Spanish users might need to balance between suggesting popular destinations and offering diverse, unique options.
7. **Limitations and Considerations**:
   - This analysis is based on users who reviewed at least 5 locations in an area, which might not represent all Spanish travelers.
   - We don't have information on the total number of Spanish users in the dataset, which could affect the interpretation if the sample size is small.
   - The data doesn't account for factors like age, income, or specific regions within Spain, which could provide more nuanced insights.

In conclusion, Spanish users in this dataset demonstrate diverse travel patterns when considering the overall set of destinations (as indicated by the low Jaccard similarity). However, they show a slightly higher level of similarity when focusing on common destinations between user pairs (as shown by the higher Overlap similarity). This suggests a travel culture that values diversity and exploration, while still having some common preferred destinations. Further investigation into specific travel preferences, popular destinations, and demographic factors could provide more detailed insights into these patterns.

2.1.10 Comparison with British, American, and Italian users

```
UNWIND ['United Kingdom', 'United States', 'Italy'] AS nationality
MATCH (u1:User {country: nationality})-[r1:review]->(a:Area_4)
WHERE r1.NB >= 5
WITH nationality, u1, COLLECT(DISTINCT a) AS areas1
MATCH (u2:User {country: nationality})-[r2:review]->(a:Area_4)
WHERE r2.NB >= 5 AND u1.id < u2.id
WITH nationality, u1, u2, areas1, COLLECT(DISTINCT a) AS areas2
WHERE SIZE(areas1) >= 5 AND SIZE(areas2) >= 5
WITH nationality, u1, u2, areas1, areas2,
     [x IN areas1 WHERE x IN areas2] AS intersection
WITH nationality,
     SIZE(intersection)*1.0 / SIZE(areas1 + [x IN areas2 WHERE NOT x IN areas2]) AS jaccard,
     SIZE(intersection)*1.0 / CASE WHEN SIZE(areas1) < SIZE(areas2) THEN SIZE(areas1) ELSE
SIZE(areas2) END AS overlap
WITH nationality, AVG(jaccard) AS avgJaccard, AVG(overlap) AS avgOverlap, COUNT(*) AS
pairCount
RETURN nationality, avgJaccard, avgOverlap, pairCount
ORDER BY avgJaccard DESC
```

Results:

Nationality	Average Jaccard	Average Overlap	Pair Count
United States	0.0785	0.1099	1225
Italy	0.0490	0.0711	1225
United Kingdom	0.0321	0.0373	903

Explanation: Explanation of differences between nationalities:

1. **Similarity Measures**:
    - The Average Jaccard and Average Overlap scores are consistently higher for American users, followed by Italian users, and then British users.
    - This suggests that American users tend to have more similar travel patterns among themselves compared to Italian and British users.
2. **Pair Count**:
    - American and Italian users have the same number of pairs (1225), while British users have fewer pairs (903).
    - This could indicate that there are more American and Italian users in the dataset who visited at least 5 distinct areas, or that these users have more diverse travel patterns leading to more unique pairings.
3. **Differences Between Nationalities**:

    a) **United States**:

    - Highest similarity scores (Jaccard: 0.0785, Overlap: 0.1099)
    - This suggests that American users in the dataset tend to visit more similar areas compared to the other nationalities.
    - Possible explanations: More homogeneous travel preferences, concentration on popular tourist destinations, or potentially a larger sample size leading to more diverse pairings.

    b) **Italy**:

    - Middle-range similarity scores (Jaccard: 0.0490, Overlap: 0.0711)
    - Italian users show moderate similarity in their travel patterns, less than Americans but more than British users.
    - This could indicate a balance between visiting popular destinations and exploring more diverse locations.

    c) **United Kingdom**:

    - Lowest similarity scores (Jaccard: 0.0321, Overlap: 0.0373)
    - British users demonstrate the least similarity in their travel patterns among the three nationalities.
    - This might suggest more diverse travel preferences among British users, or a tendency to explore less common destinations.

4. **Interpretation of Measures**:

    - The Jaccard similarity is consistently lower than the Overlap similarity for all nationalities. This is expected as Jaccard considers the union of all areas visited, while Overlap focuses on the smaller set.
    - The difference between Jaccard and Overlap is most pronounced for American users, suggesting they might have a larger variety in the total number of areas visited.

5. **Possible Explanations for Differences**:

    - Cultural factors: Different travel cultures and preferences among the nationalities.
    - Geographic factors: The location of the home country might influence travel patterns (e.g., Americans might focus more on certain regions due to distance).
    - Economic factors: Differences in average vacation time or travel budgets could affect the diversity of destinations.
    - Tourism marketing: Certain destinations might be marketed more heavily in some countries than others.

6. **Limitations**:

    - This analysis only includes users who reviewed at least 5 locations in an area, which might not represent the entire user base.
    - The data doesn't account for the total number of users from each country, which could skew the results if the sample sizes are significantly different.

In conclusion, while American users show the highest similarity in travel patterns, followed by Italians and then British users, the overall similarity scores are relatively low (all under 0.11). This suggests that even within nationalities, there's significant diversity in travel preferences and destinations visited. Further investigation into the specific areas visited and other demographic factors could provide more insights into these travel patterns.

Conclusion

This analysis of bi-partite graphs using Neo4j has provided valuable insights into user behavior, travel patterns, and similarities among users from different nationalities. Here are the key findings and conclusions:

1. **User Engagement**: We identified highly active users, such as the top French reviewers (User 70 and 76), who contributed significantly to the platform with thousands of reviews. This highlights the importance of "super users" in generating content and potentially influencing overall ratings.

2. **Travel Diversity**: Users who visited the most distinct areas (e.g., Users 2639 and 387312) demonstrated more diverse travel patterns compared to those who simply reviewed frequently. This suggests that the platform benefits from both depth (frequent reviewers) and breadth (diverse travelers) of content.

3. **Similarity Measures**: Various similarity measures (Jaccard, Overlap, Euclidean, Cosine) were used to compare user behaviors:

    - Jaccard and Overlap similarities consistently showed low to moderate values, indicating diverse travel patterns even among users from the same country.
    - Euclidean distances and Cosine similarities based on ratings showed high similarities, suggesting consistent rating behaviors across different user types.

4. **Rating Behavior**: Despite differences in travel patterns, users generally showed high consistency in their rating behaviors. This suggests that the platform's rating system is being used consistently across different user types, which is valuable for recommendation systems and understanding user preferences.

5. **National Differences**: Comparing users from different countries (Spain, United States, Italy, United Kingdom) revealed interesting patterns:

    - American users showed the highest similarity in travel patterns, followed by Italian and then British users.
    - Spanish users demonstrated diverse travel patterns with low average similarities.

   - These differences highlight the importance of considering cultural and geographic factors in understanding travel behaviors.

6. **Implications for the Platform**:

   - The diversity in user behavior (frequent reviewers vs. diverse travelers) contributes to the richness of the platform's content.
   - High similarities in rating patterns suggest good data quality and reliable user input.
   - The platform benefits from having users with varied travel experiences, providing both focused, in-depth coverage and broad, diverse perspectives.

7. **Future Directions**: This analysis opens up several avenues for further investigation:

   - Examining the specific areas visited by different user groups to identify popular destinations or emerging trends.
   - Analyzing temporal patterns in user behavior and travel preferences.
   - Investigating the impact of demographic factors (age, income, etc.) on travel patterns and rating behaviors.
   - Developing more sophisticated recommendation systems that balance between suggesting popular destinations and offering diverse, unique options.

In conclusion, this graph-based analysis has provided deep insights into user behavior on the travel platform. It highlights the value of diverse user contributions, the consistency in rating behaviors, and the importance of considering cultural differences in travel patterns. These findings can inform strategies for user engagement, content curation, and personalized recommendations, ultimately enhancing the platform's value for travelers worldwide.