Increasing accuracy in map visualizations of tagged networks

Tazmilur Saad & Tuan Wang

Outline

- Introduction
    - Motivation
    - GMAP
    - Limitations of GMAP
    - Formalizing the problem
- Similarity
    - Notions of similarity
    - Calculating similarity
    - Modifications
- Miscellaneous
    - Maintaining aesthetics
    - Algorithmic complexity
- Conclusion

Motivation

- Suppose that you have a dataset D where each individual record i has a set of tags.
- How can we best visualize this network?


```
"media": [
  {
    "id": 30002,
    "title": {
      "userPreferred": "Berserk"
    },
    "popularity": 93568,
    "genres": [
      "Action",
      "Adventure",
      "Drama",
      "Fantasy",
      "Horror",
      "Psychological"
    ],
    "averageScore": 92
  },
  {
    "id": 30013,
    "title": {
      "userPreferred": "ONE PIECE"
    },
    "popularity": 89948,
    "genres": [
      "Action",
      "Adventure",
      "Comedy",
      "Fantasy"
    ],
    "averageScore": 91
```


Fig 1. Anilist manga sorted by score

GMAP

- Gansner, Hu & Kuborov (2010) published GMAP – an efficient algorithm for visualizing networks as maps.
- GMAP works by:
  - Applying a force directed algorithm to embed the graph
  - Performing cluster analysis
  - Generating Voronoi diagrams and adding noise to the boundaries
  - And finally, coloring the map

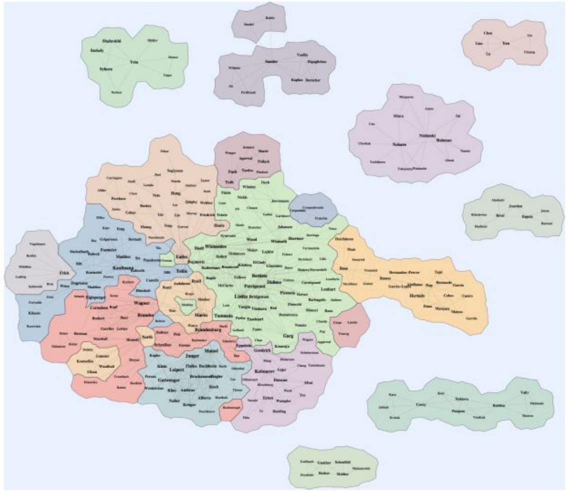

Fig 2. Application of GMAP

Limitations of GMAP

- False positives
  - Connected components end up next to each other without a strong connection between them.
  - Implies association between elements of the components
- False negatives:
  - Regions are not next to each other but they should be.
  - A harder problem to solve on a 2D-plane.

Formalizing the problem

- Assume that we are given a graph $G(V,E)$ and a mapping $\mathbf{\Gamma : V \rightarrow T}$ from the vertices to a set of possible tags. An edge exists between vertices i and j if and only if $| \Gamma(i) \cap \Gamma(j) | \geq 1$.
- We need to output a visualization V of the graph G such that the number of dissimilar regions are minimized.

Notions of Similarity

- How do we define the concept of "similarity" between (sub)graphs or the elements within?
  - Isomorphism, edit distance, common subgraphs, etc
- Intuitive notion of similarity between regions
  - Vertex similarity as a local property
  - Jaccard similarity, cosine similarity, etc.
  - Holmes et al (2005)'s definition of similarity
    - ***i*** is similar to ***j*** if ***i*** has a neighbor ***k*** that is also similar to ***j***
    - A recursive definition

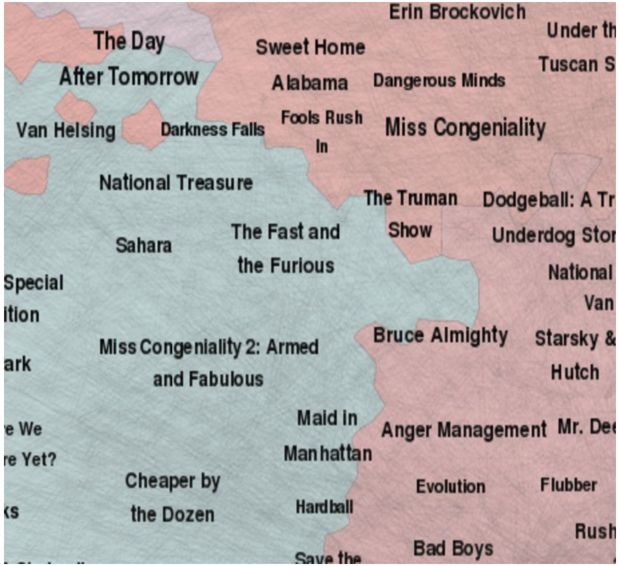

Fig 3. A map of movies

Calculating similarity

- A mathematically complicated process.
- $S_{ij} = \phi \sum_v A_{iv} S_{vj} + \psi \delta_{ij},$
- A vertex is similar to itself.
- For large graphs, eigenvalues provide approximate number of expected paths.
- Finally similarity matrix can be calculated as:

$$\mathbf{S} = 2m\lambda_1 \mathbf{D}^{-1} \left(\mathbf{I} - \frac{\alpha}{\lambda_1} \mathbf{A} \right)^{-1} \mathbf{D}^{-1},$$

$$\mathbf{DSD} = \frac{\alpha}{\lambda_1} \mathbf{A}(\mathbf{DSD}) + \mathbf{I}.$$

Modifications

- Holmes et al’s algorithm was designed for adjacency matrices of only 1s and 0s.
- Since we are using tagged networks, we will utilize Γ to modify the adjacency matrix to operate on values in [0,1].
- We do not need to work on the entire graph – only on connected components.

Maintaining Aesthetics

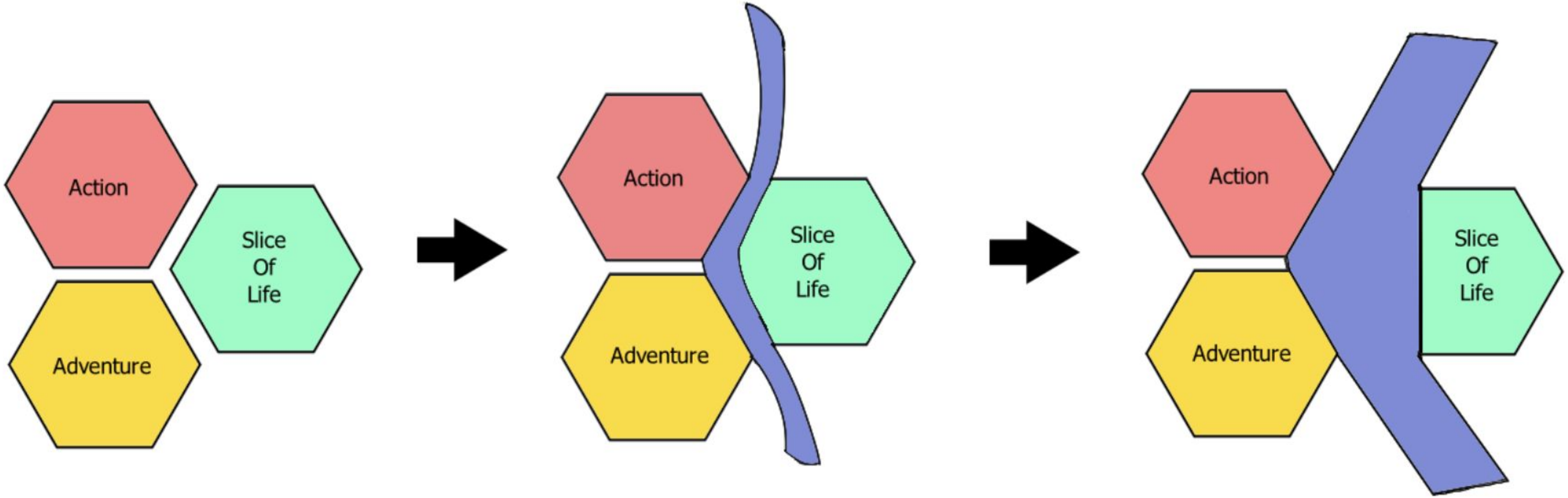

Fig 4. Adding rivers/lakes.

Algorithmic Complexity

- Complexity of GMap is O(|V|log|V|).
- Addition of extra components does not affect asymptotic bound.
- Calculating similarity uses only linear algebra operations
  - Matrix multiplication is typically O(|V|^(2.37))
  - Generally much faster and hardware optimized
  - Lots of libraries exist even for high level languages like Python

Conclusion

- We have shown a possible way to improve the aesthetics and accuracy of GMap by embedding a bounded number of points and adding repulsive force from those points.
- The location of these points can be found by calculating the similarity matrix for a subset of the points in the graph.
- Calculating similarity is expensive asymptotically but generally fast.
- Further work can be done to find a way to reduce the number of false negatives.

References

1. Gansner, Emden & Hu, Yifan & Kobourov, Stephen. (2010). GMap: Visualizing graphs and clusters as maps. Computer Graphics and Applications, IEEE. 30. 201 - 208. 10.1109/PACIFICVIS.2010.5429590.
2. Leicht, EA & Holme, Petter & Newman, M. (2006). Vertex similarity in networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 73. 026120. 10.1103/PhysRevE.73.026120.
3. Zager, Laura & Verghese, George. (2008). Graph similarity scoring and matching. Applied Mathematics Letters. 21. 86-94. 10.1016/j.aml.2007.01.006.