Dynamic Link Prediction

Rumeet Goradia (rug5), Rohan Shah (ras513)

Rutgers University: CS550 — Massive Data Mining

Requirements

The codebase is implemented in Python 3.8.5. Package versions used for development are specified below.

```
networkx          2.5
tqdm              4.50.2
numpy             1.19.2
```

Datasets

All datasets are available in the codebase in the `data` directory. They are also available in their unformatted forms on SNAP.

Usage

The following files can be run without any command-line arguments. The data they produce should already be present in the `graphs` directory and `communities` directory respectively, however, so there should be no need to run them.

- `graph.py` — generate graphs for all datasets
- `community_detection.py` — generate community sets for all datasets (also presents modularity of and number of communities in each partition)

The following files must be run with one of the arguments below as a singular command-line argument; if multiple arguments are entered, only the first one with be registered. Please note that running each of these files produces data files several gigabytes large.

- `neighbor_similarity.py` — calculate similarity between nodes based on the Jaccard coefficient of their common neighbors
  (creates `neighbor_similarity_data` directory and subdirectory for inputted dataset)
- `similarity_propagation.py` — calculate similarity between nodes based on similarity propagation of their features
  (creates `similarity_propagation_data` directory and subdirectory for inputted dataset)
- `dynamic_link_prediction.py` — calculate similarity between nodes using a combination of the 2 aforementioned methods, as described in the report
  (creates `dynamic_link_prediction_data` directory and subdirectory for inputted dataset)

Possible inputs for the above files include:

- `lastfm`
- `twitch-de`
- `twitch-en`
- `twitch-es`
- `twitch-fr`
- `twitch-pr`
- `twitch-ru`

Notes

Convergence, as required for the completion of `similarity_propagation.py` and `dynamic_link_prediction.py`, takes a long time (especially for the `twitch-de` and `twitch-fr` datasets due to their large number of edges). Therefore, after the initial similarity matrices are calculated, these two programs can safely be stopped at any time via a `KeyboardInterrupt` (`Ctrl+C`). The most recent similarity matrix will be saved in the respective subdirectory of the inputted dataset under the respective directory of the program that was run. To resume the iterative process of calculating the similarity matrix until convergence, simply delete the `results.txt` file in this location and re-run the program with the same command line argument.