**A “People You May [*Want to*] Know” Algorithm for Connecting Individuals Based on Personal Goals**

Nicholas Christman, nc2677

B13: Human Capital Analysis – Understanding Personal Value and Objectives

14 February 2020

**Abstract:** Social networks are known to bring individuals together, connecting people, groups, and organizations. This is accomplished using various different methods and parameters to predict friends based on some criteria. Arguably, the most popular method high-profile platforms use is the “People You May Know” approach – recommending friends (or connections) based primarily on the user’s current network of friends and groups. In this project, another alternative approach is explored, suggesting “People You May **Want to** Know” based on common goals shared by users. Standard goal topics will be established, and a user’s goals will be processed using a Natural Language Processing (NLP) toolbox to classify the context of the goal, detecting the specific goal topic and classifying the user’s goal appropriately. The network will be modeled as a graph where users represent the nodes and the goal topics serve as the edges. From a high level, the expected result is a “top 5” connection recommendation list for each user – a visualization of the social network will also be developed.

1. **Literature Review:**

Although other parameters are considered [4,5], it is fairly well known that high profile social network platforms’ (such as Facebook, LinkedIn, etc.) primary method for suggesting friends is based on the users direct friends’ social network (i.e., friends-of-friends or mutual friends) [5]. This makes an effective platform for individuals seeking new friendships or romantic partners, reconnecting with past friends, etc.

During the project validity research phase, however, it was discovered that proposals for network connection suggestions based on something other than mutual friends have been documented. For instance, Phepale Pranoti, et. al., suggests recommending friend connections based on the lifestyle of the individual [2]. In this context, the user’s lifestyle is defined by the daily activities carried out and locations they visit. The proposed system uses text mining techniques to analyze users’ lifestyles and generates a tabular Friend-matching graph.

Another example for friend recommendation is based common and uncommon behavior criteria [3]. In the article, the authors propose a hybrid approach of suggesting friends not just on common behaviors, but also using the uncommon behaviors. The example used in the article states that two users who like rock music should be recommended as friends, but that the uncommon behaviors (e.g., locations, places of employment, etc.) can make friend recommendation more effective.

1. **Introduction to the Project:**

The project proposed herein extends the research performed in this field and goes beyond the “People You May Know” paradigm, which is not sufficient for someone interested in creating a network of like-minded individuals. What appears to be common in the most recent research of friend recommendation is the use of social characteristics – place of residence, work habits, personal preferences – all of which connect individuals who seek a friendship or personal connection. What makes this project unique is it seeks to generate a network of connections based on goals, allowing individuals to find like-minded groups necessary for emotional support, accountability or even business ventures – and this is regardless of their current social network. A high level example of what this model might look like is illustrated in Figure 1 below.

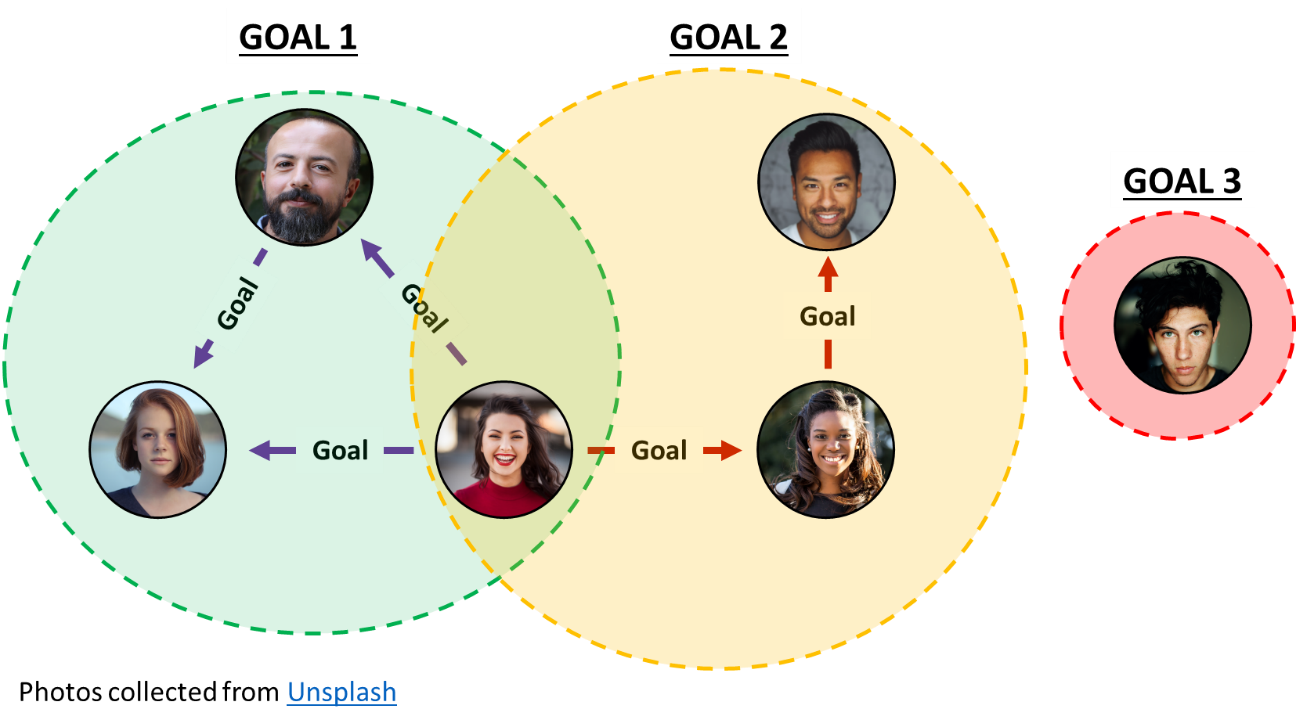
****

Figure : Example of how users might be connected by similar goals.

1. **Introduction to the Dataset:**

Identifying a complete dataset for this project has proven difficult, because user profile data combined with contextual goals does not appear to be available. This is the first and biggest challenge to overcome for this project to succeed. If a dataset is not identified by the Preliminary Analysis step in Phase 2 (see schedule below), then an alternative approach will be taken.

With that said, the data to be used for this project includes approximately 100,000 pseudo-Facebook user profiles that are available on Kaggle[[1]](#footnote-1). This data includes unique user IDs (these will be the graph ***nodes***) and other personal information such as age, gender, and date of birth. For this project to be valid, these users will need goals – these goals, if a dataset is not found, will be emulated and randomly distributed amongst the 100,000 user profiles. Although this is not ideal, it will provide the necessary goal context needed to create the links (or edges) between each user. All of this will be completed during the data preprocessing task of Phase 2.

1. **Plan:**

In Table 1 provided below, the schedule for the project is defined.

* The first phase has been to confirm the validity of the objective and perform some market research to determine if this is a novel approach to suggesting connections between individuals. The second piece to this phase was identifying a dataset (or datasets) to be used throughout this project – this task has proven difficult, but an alternative plan has been developed, as explained in the previous section. **At the end of this phase, Milestone 1 is reached and the project can continue.**
* The second phase is the highest risk – it is expected that the dataset being used will provide the level of details necessary to generate reliable edges and connect users (nodes) to each other based on similar goals. Some trial-and-error may be necessary to get the context/key words of a goal to generate useful edges (this will be evaluated in the final report). **At the end of this phase, Milestone 2 is reached and the proposed ecosystems appear to work for the given dataset.**
* The third and final phase is intended to wrap up the analysis performed in phase two and generate a visualization that effectively describes the results. This visualization will be the in the form of an undirected connected components graph, highlighting the different “goal topic” subgraphs and the users within those subgraphs. As time permits, a demonstration of friend recommendation using the live dataset will also be provided. **At the end of this phase, Milestone 3 and Milestone 4 are reached and the final project report/presentation are delivered.**

|  |  |
| --- | --- |
| Table 1:Project schedule | |
| **Phase** | **Task/Deliverable** |
| 1 | * Research datasets & write proposal |
| **MS 1** | **Project applicability verified (pending feedback)** |
| 2 | * Setup ecosystems (GCP, Neo4j), refactor/store dataset (GC bucket) * Verify dataflow (John Snow LABS NLP, Neo4j-Spark-Connector) * Data preprocessing (fusing datasets) * Preliminary analysis in GCP (Spark, NLP) * Preliminary results stored in Neo4j Graph DB |
| **MS 2** | **Connected components as expected (in Neo4j UI)** |
| 3 | * Final visualization (D3, Neo4j javascript driver) |
| **MS 3** | **Visualization gives clear picture of results** |
| **MS 4** | **Final report & presentation** |

**Reference:**

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

[2] Phepale Pranoti, & Prof. C.V Longani. (2016). A System for Suggestion of friends for Social Networks. International Journal Of Advance Research And Innovative Ideas In Education, 2(1), 287-289.

[3] Shehu, S. (2017). Friend Suggestion System for the Social Network Based on User Behavior. International Journal of Computer Science, Engineering and Information Technology, 7(5), 15–19. doi: 10.5121/ijcseit.2017.7502

[4] <https://www.linkedin.com/help/linkedin/answer/29/-people-you-may-know-feature-overview?lang=en>

[5] <https://www.facebook.com/help/163810437015615>

1. <https://www.kaggle.com/sheenabatra/facebook-data> [↑](#footnote-ref-1)