Entity Resolution between Online Social Networks
A Comparison of the Professional versus the Social Identity

The goal of this research to identify and associate the identifiable attributes of a real-life person (i.e., entity) to a social profile found on a professional network, such as LinkedIn. We then want correlate the LinkedIn profile to one found on a social network, such as Twitter. The candidate profiles will be collected using alumni information from a specific institution of Higher Education. We will then compare profile attributes to determine whether a probabilistic match is sufficient to assume the same entity based on the strength and number of overlapping attributes. If we can obtain data on LinkedIn connections and Twitter followers, then we will perform a community analysis to determine the similarity of each network. The degree of network similarity will serve as an additional identifiable attribute.

To add additional context to the personally identifiable information, we might also consider examining the contents and URLs in tweets to see whether they are consistent (i.e., same theme) with the stated industry or degree program on the LinkedIn profile. It's possible that, even in a non-professional environment, a user will tweet about what they know best.

1. Data Set

First, we need to compile an initial set of candidates for each online social network based on a limited amount of preliminary data. For alumni, the identifying attributes might include the name, institution, and degree program.

Question: How do we obtain this list? Electronically, web scraping?

2. Profile Search

a. Professional Network: LinkedIn

In this step, we attempt to find all possible matching profiles using either the dedicated API or we can use a search engine to locate public profiles while restricting the domain to LinkedIn. The API will return data in a structured format, such as JSON. The search engine will return URLs which we will then need to part to determine the associated profile ID. Since real names are more likely in this environment, we will use the first and last name and the name of the institution as the query criteria.

Caveat: LinkedIn Developer's terms of use – Excerpt for member profile data:

> You may perform a one-time capture of the user's Profile Data and store that Profile Data if you **have the consent of the user** to do so, for example, so that a user applying for a job at your company can provide you with a copy of their profile. "Profile Data" means the name, photo, headline, contact information, experience, education, summary, and location of a LinkedIn member. Profile Data excludes connections, network updates, job listings, groups, companies, and any other Content. The process for obtaining member consent must meet our specifications. If you want to refresh the user's Profile Data, you must ask the user for consent before doing so. You must use stored Profile Data solely for the benefit of the LinkedIn user that

granted you permission to access it, for example, to evaluate that user for a position in your company.

Question: How do we obtain this consent for a large group of people? Known entities or an email blast with a "request to participate?"

b. Social Network: Twitter

For each identified LinkedIn profile, we will create a derivation of the profile ID to search for potential matching profiles on Twitter. Examples of the derivations might include: firstLast, first.Last, nicknameLast and so on. (Refer to Carlton Northern)

3. Data Standardization

For consistency between networks, we will need to map profile attributes between networks to ensure only comparable attributes are considered. We will also have to address known limitations in acceptable formats and data sizes.

Generic Attribute	**LinkedIn Profile**	**Twitter Profile**
ID	ID	Id_str
First name	First-name	
Last name	Last-name	
Full name	Formatted-name	name
	summary	bio
Geo data	location	location
	Contact info - websites	website
	Twitter account*	

Sampling of potential attributes

4. Matching Algorithm

The first matching algorithm will attempt to score the similarity of profiles between both social networks. The process will start with a direct comparison of similar attributes. For string attributes, we can test for fuzzy similarity using a distance algorithm (e.g., Levenshtein). For other attributes, such as a web link, a direct equality would be more appropriate. The scoring system will assign a positive score for matching attributes, while deducting points in the event of a mismatch. If the resulting similarity score of the two profiles is above some arbitrary threshold (TBD), we will consider the profiles to be a linked pair. Since there can be many such pairs, we will use the technique of locality sensitive hashing to

perform pairwise binning of the candidate pairs using a predetermined sort order. Once sorted, the pairs most likely to be matches will be closest in distance from one another.

5. Community Analysis using Graphs

To facilitate a graphical analysis, we will repeat use the first-level, direct connections from LinkedIn and the Twitter followers. We should note that LinkedIn doesn't allow you to directly access second level (friend of a friend) connections. We can use these nodes to create a set of graphs that will allow us to determine (1) whether any communities are observed (clusters) and (2) whether there is any similarity between the two networks.

6. Evaluation

The LinkedIn profile contains multiple fields for the user's Twitter account. This field can be retrieved and used to determine whether the matching algorithm has located the correct profile.

Question: Is this an acceptable evaluation scheme? Or, should the matching occur within an anonymous environment so no bias is introduced?

7. Toolkit

We will Python and any applicable third-party libraries for any web scraping and to interact with the social APIs. A graph database, such as Neo4j, will be used for the community analysis and visualization.