Social Network Outreach Analysis
Capstone 2 Project Proposal
Springboard Data Science Career
Track Eric Cruz

The proposal should address the following questions:

- What is the problem you want to solve?

Develop and test a Data Science strategy for Social Network Outreach. The primary issue is how to identify the target members on a popular selection of social media outlets such as LinkedIn, Facebook, Instagram, and Twitter. Due to increasing efforts to limit the ability to compile the data and share the respective API's , a toolkit of scraping and searching tools will be selected for the purpose of digitizing the information and automating the collection of search results. The end goal is to develop a strategy for engaging the target audience on various platforms by analyzing participation rates or other trends. For example, we may expect to observe generational trends, e.g. younger people have left Facebook or reduced usage in favor of Instagram, older people are likely to be on LinkedIn but not Facebook. The actual outreach will be conducted by individuals reaching out to their peers, using curated contact lists generated from the search phase.

- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

The client at the lowest level cluster is the Men's or Women's college tennis team, with next level groupings to combine the genders for a school's program, and group schools based on Conference and Regional affiliation. The high level client is the College Tennis Alumni Association, which encompasses all the clusters and already functions as a social media outlet with members.

The client will decide on social media outreach strategies to combine with traditional method of email which may or may not be verified contact addresses.

- What data are you using? How will you acquire the data?

The data regarding eligibility as a target member will be obtained from the team websites in conjunction with websites which report match results and recruiting information. The eligibility data will be scraped. Other data sources include membership lists from groups on LinkedIn, Facebook, and the College Tennis Alumni Network.

- Briefly outline how you'll solve this problem. Your approach may change later,

but this is a good first step to get you thinking about a method and solution.

The compilation of the eligibility list is a straightforward exercise, but the development of tools for identifying the target individual amongst several possible profiles on each social media outlet will be the the dynamic challenge, given the lack access to the API's. By using commercially available scrape and search tools, the product limitations will lead to a roadmap for a script to be developed should access be obtained at a later date.

- What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The deliverables are a set of tools for conducting the research on a list of names, and a report describing the challenges and solutions. The tools will gather information for matching on keywords and photos. A prototype campaign will be conducted on a sample team cluster, and a sample regional target list will be built. The actual outreach campaign completion is out of scope, as the problem to resolve is to develop a scalable strategy, rather than complete the execution.

- ❏ The submission demonstrates that the student has articulated a **problem statement** that is within scope of this course.
- ❏ The submission demonstrates that the student has selected a problem with a **dataset** that is relatively clean, easy to use, and within scope.
- ❏ The submission demonstrates the student is able to justify a significant **problem** to solve for a realistic client (even if imaginary).
- ❏ The submission includes a strong **justification** of how the client can utilize the **outcomes** of the project.

Problem Statement: Using data for College Tennis Alumni Network candidates, collect and label data on targeted individuals, and engage them to expand the network.

Dataset: The data will be collected from websites using scraping tools and other utilities, and may include a comparison of different methods for obtaining and storing data. There will be outreach based on lists for a local alumni organization included in the data collection exercise.

Clients: The clients are the CTAN, or College Tennis Alumni Network, and local level Cornell Men's and Women's Tennis Alumni Associations.

Approach: At the local alumni level, there is a list compiled over 4 years, recording attempts to contact via email for a summer event. A fair number have never received a response. The list needs to be reviewed and updated with the current coaches, as the Women's coach has been replaced. The 50th Anniversary of Women's Tennis at Cornell will honor Hall of Fame inductees on campus in June 2020. At the CTAN level, the growth strategy will rely on member volunteer committees at the local level, and I will participate in NYC chapter of CTAN, creating opportunities to create outreach campaigns and record the data. Names and College affiliation, including dates played, from the parent company ITA scoreboard online, will be the baseline for membership eligibility and matching will be attempted on social media networks given scraping capabilities from each source. Natural Language Processing techniques will be used to differentiate between many people with the same or similar name, using school names and tennis among other cross reference searches. Sample target collection data includes whether person is found on Facebook, LinkedIn, etc., and whether school website identifies as "Captain". Outreach will be targeted toward captains, asking for their participation expanding the network amongst their peers and beyond if accessible. Data will be collected on network influence statistics.

The Initial Contact List Data

Members of the College Tennis Alumni Network with a NY address

The CTAN was launched in April 2018. It started hosting events at the National Indoor Team competition. It was advertised as a reception for all College Tennis Alumni, with no obligation to join the organization, and no charge to the alum. We can assume that the people attending the event were there to support their teams or the event anyway, and that many of them are the target members, meaning former players of those teams or their rivals. The organization did not have a specific outreach strategy, other than word of mouth and advertising on their own website which publishes the official results for all of College Tennis. Other events were planned as networking events for Alumni from all schools, held at venues such as a restaurant or bar, not connected with a specific tennis event. In August 2019, an event was planned in New York which was advertised as a pre-US Open gathering for tennis enthusiasts from all schools. Starting in 2020, the CTAN defined a membership growth strategy of setting up local chapters in cities such as Chicago and New York, run by volunteers. As the New York chapter is being formed, we are starting with the list of people who signed up for membership in CTAN, with an address in NY state.

The distribution of people by graduation year, school, and gender is surprisingly diverse. With 142 people, 88 schools are represented, and members come from 36 different graduation years in the 46-year range between 1978 and 2024. The gender distribution is 50/50.

Cornell Women's Tennis Alumni

The team Alumni list has been manually curated for the purpose of outreach for the 2020 event. Starting with the contact list provided by the school's Alumni Affairs representative, the roster list generated from the school's database was compared to the archived pdf files going back to the program's start in 1971. In other words, the lists were typed in manually from images of paper copies for the comparison. The Alumni Association for the Cornell Men's and Women's teams became active about 5 year ago, in terms of hosting events and conducting outreach to support them. In addition, there has always been a supporter list maintained by Alumni Affairs, and an annual phone-a-thon calling for the current players to solicit donations from their assigned leads. The lead list identifies whether the target is a team alum, so expectations are that supportive team alums have already been identified. However, this can work in reverse in the sense that less than enthusiastic supporters may have opted out of being contacted, or otherwise learned how to avoid being solicited for donations. For example, many of the email addresses appear to be suspicious on the surface, such as those with a cornell.edu extension for someone who has already graduated. Other extensions such a aol.com seem suspicious just because they seem obsolete, and others which are now commonplace may be used specifically for the purpose of maintaining addresses which are low priority for responsiveness. The phone-a-thon is based on phone numbers, so the analogy to email addresses is on a slightly different level.

As part of the Alumni revival effort, we started hosting a summer event in the NYC area, which involves playing tennis followed by a BBQ dinner. The event is for the combined Men's and Women's alumni associations, attended by the coaches, current and former players and their guests. As part of the outreach effort, I started using the email contact lists from Alumni Affairs as well as the Coaches' versions they maintain for sending out newsletters, to invite people from the list based on location information. Namely, if the address is in a state close enough to drive, I would prioritize those people, and expand outward in later cycles. I added a label for "ROSTER LIST" and "OutreachDate" if applicable. The Women's contact list from 2018 included both team alumni and supporters, and it was clear from the previous 2 years that many on the list were non-responsive, not to my local outreach effort, but to the Coach according to verbal account. The list I received for 2019 was intended to be alumni players according to the database query. There were 61 new names but approximately one-third did not have a email address. The curated list for 2020 has the remaining 18 names from the past list, plus the 14 new names from current and last year, bringing the total to 271 players.

The Data Download

When scraping the rosters from college websites, the fields available for collection usually include Name, Year in School, Home Town, High School, and Roster Year is attached to each download. The more recent years include headshot photos, but the scraping method used did not pick up those links in most cases. The scraping method also did not pick up the link to the profile in many cases as well, and these were included when choosing the settings with the scraping tool. This could be due to the settings on the websites, or capacity of the scraping tool, and will be investigated further. The availability of rosters varies by school with respect to how far they go back. The initial sample indicates that many go back to mid-2000's and others

another 20 years or so. Very few have a full history, but many programs maintain an All-time roster list which can be downloaded in pdf format.