Abstract:
- Analyze the connectivity of an online social network, LiveJournal
- Use this analysis to create a decentralized searching algorithm which can traverse the network to find targets from random sources

Small-World Graphs:
- A recently categorized phenomenon in graph theory, small-world graphs exemplify the “six degrees of separation” theory in traditional social networks
- By definition, a graph is considered a small-world graph if it has a low diameter, generally on the order of \(\log(n)\), and there exists a high probability of finding a target from a source using only local information, and traversing a small number of contacts along the way.
- There already exist many ways to find the diameter of a graph, but ensuring a short path from random sources to targets using a decentralized search is a much more versatile problem

LiveJournal:
- The social network that was analyzed for this project was LiveJournal, a site which enables users to create their personal online journal.
- Each user has a profile, where they can list information such as their location, interests, schools, as well as friends who are other LiveJournal users.
- The site has approximately thirteen million users.

Retrieving the Data:
- Acquiring information online, in this case HTML source code, is very expensive from a time constraint perspective.
- Gaining a lot of data at once in order to run processes on it offline is ideal.
- Using a regular expression matcher to identify specific patterns in the LiveJournal source code for a user’s profile page, it is possible to distinguish and maintain any information about the user offline.
- Treating the site as a graph where user profiles are vertexes and the listing of another user as a friend constitutes an edge, a breadth-first search centered at one vertex yielded information on almost 200,000 users.
- The sample had a diameter of ten, and went five steps out from the source during the breadth-first search.

Analysis:
- Tens of thousands of traversals of the graph were done, attempting to give insight into the connectivity of the graph, specifically what would make a short path likely between two vertexes.
- Many tests involved finding the average path length for users who shared multiple interests, communities, the same geographical location, and other matching pieces of information that users could list.
- Searches were done for possible hubs, central connecting points of the graph through which a high number of short paths exist.

Decentralized Search:
- For the decentralized search, I decided to implement a greedy algorithm, one which always extends the path computed to the vertex “closest” to the target.
- The distance of a vertex to its target was completely dependent on the information gleaned from the network in previous analysis.
- Generally, the more items in common a vertex had with the target, the “closer” it is. The more significant the correlation in the original results, the shorter the perceived distance is between them.
- This would be run on the “real” network, traversing the actual web pages of LiveJournal as it ran.
- The algorithm would only have access to information it sees along the way, it begins only with the data for the source and the target.

Fine-tuning the Search:
- The first version of the search, which used explicit values from the data, did not perform well.
- A reevaluation of the changes that occur during a transition from a smaller sample of the graph to the entire network was necessary.
- Among the most important changes involved highly favoring users who listed the same location.
- It is likely that the sample focused on some geographical location rather than some specific set of interests or communities.
- Previous research has indicated strong correlations between short paths and close physical proximity in social networks.
- Therefore, the average path length found in the sample for geographical location is representative of the entire graph, whereas the average path length for other items, such as interests, went up with the size of the network.
- By favoring location first, the search was able to home in on the target, and then focus on secondary characteristics such as shared schools.

Final Results:
- With the new heuristics for distance, the decentralized search is successful with high probability.
- Given a reasonable amount of information listed by the target, the search computes a path from the source to the target of about twenty to thirty in length.