# CS 6630 Project Proposal
# Visualizing SIGGRAPH Publications

Kui, Wu and Hoang, Duong
TODO (u0930062)
TODO@TODO u0933062@utah.edu

October 23, 2015

## 1 Introduction and background

When surveying a new field, researchers often want an overview look at the important papers in the field, how they are related, and how popular trends in topics or techniques have been evolving over the years. Traditional search engines such as Google Scholar and digital libraries such as the ACM's can alleviate the task, but only when the researcher has a good idea of what he should look for. Even then, the information obtained from those tools at any given moment is not only too specific but also presented in a non-visual form, making it difficult for the user to see a bigger picture. Being both interested in doing research in computer graphics, in this project we aim to develop a tool to visualize connections between papers published over the years in SIGGRAPH, the most important conference in computer graphics. The tool should show in one place: the citation relationships among SIGGRAPH papers, important papers in each sub-field, prolific authors and their collaboration patterns, popular topics and methods in recent years, and active research institutions in each sub-field. It is our hope that such a tool can be particularly helpful to someone who wants to survey a field and summarize major results. It could also help new researchers in finding not only interesting problems to work on but also pointers to relevant publications for reference. Lastly, prospective graduate students can use information provided by the tool to make more informed decision in their application process.

## 2 Project Objectives

Our project aims to answer the following questions:

- What papers are influential in the field, in terms of citation count?
- Given a paper of interest, what papers does it cite? What are the papers citing it?

- What topics and techniques are popular, and in which time periods?
- Given a set of keywords, what are the most relevant papers/authors/institutions to these keywords?
- Who does each author collaborate with the most?
- What institutions are more active in a given field, in terms of publication count?

Answering the above questions will put a make a researcher more informed about what papers/authors/topics/techniques to pay more attention to, thereby saving time in the beginning of a survey task. It can also reveal interesting patterns and trends in the past that can help him or her better decide on future research directions for a particular topic of interest.

# 3 Data

Our data comes from multiple sources.

- Paper texts are downloaded from the ACM Digital Library. From these we extract the title, authors and their affiliations, references, and important keywords.
- BibTeX information for all papers is downloaded from the Digital Bibliographic Library Browser.
- Citation count information is queried from Google Scholar.

# 4 Data Processing

We expect our raw data to require a substantial cleanup process. So far we have collected all the SIGGRAPH papers's pdf files from 2002 to 2015, totaling 17GB of data. These pdf files will be converted to text format, from which we will extract from each paper its title, authors and their affiliations, references and keywords. The extraction process is done using a combination of homemade Python scripts and opensource tools such as ParsCit.

Meaningful keyword extraction is the most difficult step of the extraction process. Most papers provide a set of "keywords" after the abstract but in our experience, these keywords are not meaningful enough: in any given year, most keywords are used by only one paper. We are looking into using an auto-summarizing tools for this step. Because of the uncertainty of keyword extraction, all features that require keywords will be optional.

We have done cursory testing of our tools and they seemed to be able to extract the required information with high accuracy, thanks to the uniformity in format of the papers. There is still some chance that noise will show up in the extracted data, in which case some post-process cleaning has to be done. We will finally need to remove all non-SIGGRAPH

papers from the list of references to make our data more self-contained and easier to work with.

Since our data is highly relational, we plan to store it in an SQL database and query the data with Javascript. At the moment sql.js and node-sqlite3 seem to be promising candidates. Tentatively, there will be different tables, one for each of the following: papers, authors, institutions, keywords. Each entry in a table will have necessary fields, for example a paper will contain links to its authors and keywords, an author entry will have links back to his or her papers, and (optionally) links to the institution where he/she worked when a paper was published.

An alternative design is to flatten all the SQL tables into JSON files and load those instead. The SQL approach allows us to not having to load everything into memory at once, perhaps at the cost of more processing time each time users change their query/filter of the data. It is not yet clear to us how big the data will be, and which one of the two approaches will work better.

# 5 Visualization Design

# 6 Features

## 6.1 Must-Have Features

## 6.2 Optional Features

# 7 Project Schedule