***Abstract*—In this report we explore ways to automatically analyze ICLS and CSCL papers to gain an understanding of the community formed by conference participants and of the research being conducted within the community. After describing how to extract data from the original dataset we derive insights into the community using network analysis, natural language processing and text mining. We discover patterns of collaboration between participants, their migration and the global spread of the community. We extract keywords and find clusters of referenced papers and documents within the conferences.**

I. Introduction

The starting point of the project are the proceedings of the International Conference of the Learning Sciences (ICLS), respectively the International Conference on Computer-Supported Collaborative Learning (CSCL). The two conferences are organized by the International Society of the Learning Sciences (ISLS) and held in biennial alternation with each other. At the time of writing the dataset contains 4 years of the proceedings, staring from 2015 to 2018. While ICLS covers the entire field of the learning sciences, CSCL focuses on learning through collaboration with the help of communication technologies [1]. To understand the community formed by the participants of the conferences as well as to understand the research being conducted within the community, we explore methods to bring insight into the following aspects:

- The institutions and countries dominant in the conferences
- The differences between ICLS and CSCL with respect to citations, contributors and publications
- First insight into changes of the years with respect to citations, contributors and publications
- Collaboration network patterns with respect to countries, institutions and authors
- The migration of participants across institutions
- Trends of keywords over the years
- Trends of citations and evolution of the number of documents from 1977 until 2018
- Clusters of the cited documents
- Clusters of the papers written within the conferences

To this means we first present an extensive pipeline to process the textual data and extract features relevant to bibliographic analysis from the textual data. After this we perform a first bibliographic analysis based on the extracted data.

II. Data Processing

The provided raw dataset consists of papers in *pdf* format, as well as associated metadata in *xml* format. To extract information from the dataset we first convert the data to a format that is more adapt to the task of data analysis and more compatible with a variety of external libraries. We store all the resulting data in *csv* format, as csv is a human readable format and can easily be processed by all the data analysis tools we employ.

A. Processing the Metadata

We extract the metadata from the xml files using the *lxmltree* python library. We associate the contents of each xml tag to a column in csv file. As the xml files do not exactly follow the xml standard, we handle errors by iteratively parsing the xml tree. For some files, the section containing the keywords is malformed which can not be handled by the parser. This is remedied by using regular expressions to match the section contains the keywords. To each file we associate the name of the source file and the order in which an author name occurs in the xml file. Each author is associated one line in the XML file. From the associated citation string we add to the first part of the citation containing the author names, year and title. which we use to identify references in other papers to that paper. Then we add a shortened version of author names (i.e. West, R.) by considering the author order. Note that the short name extracted from the citation and the corresponding long name (Robert West) in the metadata do at times not match up. Thus the author order listed in the citation string does not always correspond to the author order listed in the metadata. After extraction, we clean the data by unifying names, as slight spelling differences exist within the same author name (i.e. Robert West, Robert A. West). When unifying the names we need to make sure to not overmerge by labeling to different authors as having the same name. First, we use a strict condition to catch small differences in names, such as an extra middle name. This is done by splitting up names into substrings based on whitespace and commas. For two names in the dataset we check whether the intersection of their sets of substrings contains at least 2 names of length at least 3. If the condition is met we consider these two names to be identical, and add the mapping from one name to the other to a dictionary. Once all tuples of names (s.t. the first name appears before the second one in order) are checked and added to the dictionary we unify the names. Using this method we reduce the number of authors in the dataset from 1951 down to 1879. Further, to merge names with misspellings or people that use nicknames (Christian vs Chris) and avoid over-merging we consider only names of authors that have similar collaboration patterns. To do this we build the graph of co-authors and check names of authors that are in the same neighborhood of a node on the graph (see methodology for further details). Then we use the *difflib* library to check for similarity of two strings and consider a name to belong to the same person if they reach the threshold of 0.8. Again names considered to belong to the same person are added to the mapping dictionary which is then applied to all names. This method again reduces the number of authors in the dataset from 1879 to 1951.

B. Handling of pdf data

To extract information from the papers themselves we need to first extract the text from the pdfs. Various parsers performing this task exist. However, parsers such as pdfminer, PyPDF2 that claim to be able to extract the text as well as metadata from papers have not been kept up to date, which leads to issues with dependencies. We found that the most reliable parser is *poppler* [2] which can be installed and then used via the command line to extract the text contained in the pdfs to *txt* format. All methods describing how to extract data from the papers work on the extracted *txt* files. The characters contained in the *txt* files are not all encoded uniformly, which means that characters that look the same to humans may be encoded in different way . Hence, for many applications we need to unify the formating. This is done using the *unicodedata* module and setting converting all characters to ‘Normal Form Composed’ form .

A convenient property of papers is that they are structured fairly regularly. The ISLS Author Guidelines [3] give us guidance on how the papers in our dataset should be structured. We use this to our advantage when extracting information from the paper text. While the vast majority of papers in the dataset follow these guidelines, exceptions do exist. We thus make sure to handle deviations from the structure.

C. Extracting the papers cited by a paper

We know that each valid paper contains a reference section at the end listing all the references cited in the paper. While tools are being developed to extract the reference section directly from pdfs, notably Scholarcy [4], they did not produce satisfactory results. That is they had issues splitting the references correctly and categorizing parts of the reference accurately. This may occur because they are trained to extract any reference format, and thus can not make many assumptions on the structure. But in our dataset papers following the guidelines should list their references in APA format. Hence using regular expressions and background on how APA references gives us better results than out of the box methods.

Regular expressions in python: Python has two modules that implement regular expressions *re* and *regex*. We used regex as –unlike the re module– this module allows to handle strings with non-ascii characters by having unicode expressions. As many authors have names with non-ascii characters this is an important capability.

To get the reference section for of a paper we use a regular expressions to find the reference header and determine the sections end by checking for acknowledgement or appendix sections in the text file previously extracted using poppler. Then, to extract individual references we split the section into a list of substrings at each new line character and glue substrings back together if they satisfy a set of conditions. We found this method to be robust as it makes sure that all properly formatted references are separated.

To unify substrings we exploit the fact that that APA references all have the same initial structure: *Lastname F.M. (when published) Title which may contain colons and other non-alphabetic chars. Where published. Who published* [5]. We thus check that the beginning of each reference contains “Lastname F.M. (when published)”. It is relatively easy to check for this using regular expressions as author names and year have a given structure. If a string does not contain such a section, then it must be part of an other reference, which is lies before in the list of reference substrings. Note that this beginning is also what is added to the metadata as identifying string. We found that in practical application it is important to first merge lines if they were clearly cut off, as papers with many authors may have the beginning on multiple lines. We can find split up lines by checking for lines starting with non-alphabetic characters or starting in a sentence. We merge these with the line above. We can identify lines by their ending as being split up, such as if they end in a comma, colon or only contain author names. We merge these lines with the lines below them. Only then can we check whether a line contains a valid citation start. If it does not we merge it with the line above it. After this merging process we validate the returned reference split by looking at the overall length and the initials, as well as checking that a random sample returns satisfactory splitting accuracy. We find that the above method of extracting the references produces good results for proper APA citations and on APA citations indeed outperforms Scolarcy when it comes to splitting up the lines. *To then get the subcomponents of a reference* is rather trivial, as we can again use the structure of APA citations. The string before the first parenthesis contains

Susan A. Yoon, Jessica Koehler-Yom, Emma Anderson, and Chad Evans
yoonsa@upenn.edu, jkoehl@gse.upenn.edu, emmaa@gse.upenn.edu, echad@sas.upenn.edu
University of Pennsylvania

Figure 1: Different formats used to reference institution

Susan A. Yoon, University of Pennsylvania, yoonsa@upenn.edu
Cindy Hmelo-Silver, Indiana University, chmelosi@indiana.edu

Figure 2: Different formats used to reference institution

the authors names, inside the parentheses we find the year or some other reference to publication time, after which the title follows until the first dot. Then a reference to the publishing venue follows.

D. Extracting Institution and Location

Now we consider the section before the abstract, which contains the authors names, their email and the institution they are affiliated with. Affiliation with institution can give us a lot of insight where people are working and gives use a way estimate where they are from. To go from this section to an exact affiliation with an institution a multifaceted approach was used.

1) Approach 1: The most robust way to get a persons institution affiliation is to use their email. Emails can easily be extracted from the header section using a regular expression and given a domain-university mapping [6] give us an onto mapping from domain to institution. The additional advantage of this method is that using the additional metadata associated to each university in the mapping, we also access information about the country of the university.

2) Approach 2: The header section itself contains information on the institution, but as different was to list authors are accepted, notably two ways as can be seen in Figure 1 and Figure 2. Hence, we first detect the structure of the header section before extracting the association. We tag the substring present in this section using a combination of *named entity detection*, using the implementation provided by *spacy*, simple string comparison matching of the names of authors and institutions we extracted from the metadata respectively from Approach 1 and tagging sections containing "emails" using regex , to detect where in the header the email is. Then, taking the tagged section, we detect which subsections correspond to which author-affiliation listing type, and extract the author email - institution mapping accordingly.

Approach 2 allows to match many more institutions to authors than Approach 1 as not every institution is present in the mapping –no companies will be in the mapping– and many participants use gmail and similar providers as their official correspondence email. However, Approach 2 may contain differences in spelling in the institution name, which need to be merged. Thus we combine both approaches, and use the corresponding institutions in Approach 1 to unify the institutions obtained in Approach 2. Furthermore, we later merge names by considering the location of the institutions.

But first as with Approach 2 institution affiliation does not allow us to automatically associate a country to the institution, we need to get this through other means. A first step is to use the top level domains present in the emails to associate a country of origin to the participants and their associated institutions. Then, to plot the community we augment the dataset with geolocation information. As a further benefit, having this information we can then easily associate a country to the institution for the emails that do not have a top level domain indicating location and merge the names of institutions where necessary based on location.

As there are a variety of way in which the name of an institution could be written –consider EPFL, EPF Lausanne– crawling the english, french and german version of Wikipedia (not using the API, or the API wrapper, as this does not give access to the desired information) was a quick and accurate way to get the location. The advantage of sending requests over using the API is due to the website redirecting to the relevant article on the institution even when misspellings are present. Requests were sent using *urllib3* and *beautifulsoup* was used to parse them get the coordinates present in the article. Regretfully not all institutions, even well known ones such as Tokyo University, have their coordinates listed on their Wikipedia. To get the remaining locations we use the geopy, which is a client for geocoding web services. These results however are more error prone and will always return a result. It is thus recommended to verify the correctness of the results. To get the country we can then again use geopy. The resulting data containing paper identifier, the authors email, its associated domain, order in which the author was cited, the institution, country, Longitude and Latitude are saved to a csv file. We use the paper identifier and author order to associate this data with the metadata to link information about authors with the papers they authored.

E. Keywords extraction

1) Text preprocessing: From the pdf parser some of the documents were corrupted with some weird caracters such as , white space strings

- We use ***beautiful soup*** and ***string*** libraries to only keep the text that we needed without any special character.

- we apply ***regular expression*** to remove numbers because we thought that numbers were not that important.
- We put all the words to lower cases
- We remove the words that were connecting parts of a sentence rather than showing subjects, objects or intent. For us these words were not that important and were only going to increase the bag of words length
- In some methods we use ***Lemmatization*** and ***stemming*** and in other methods we don't because we wanted to do a mapping with ***GloVe*** vocabulary which we will explain later on.

Lemmatization is a word variations to the root of the word (e.g. working, works, worked changed to work). ***Stemming*** is the process of reducing inflected and derived words to their word stem, base or root form—generally a written word form As the use of the stop words to reduce the BOW (bag of words size we also used lemmatization and stemming to be able to reduce the number of words without taking off any information)

Lemmatization and different form of ***Stemming*** were only used with ***TF-IDF*** while doing the Keywords extraction, otherwise we prefer keeping the words as they are because we are doing a mapping with a pretrained stanford word embedding and we do not want to loose interpretability.

For the Keywords extraction we use three methods :

The first one whe is a trivial one is to search in the pdf itself for keywords and extract them, this method is the most accurate but we found out that for 492 papers we had no keywords included. So we deciced to use two methods for keywords extraction. The first one is ***TF-IDF*** and the second one is ***Rake***.

- ***TF-IDF***: Tf–idf is one of the most popular term-weighting schemes today, in information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general [7]

With TF-IDF we sorted for each document the words with the highest score and then selected the the 7 first words for each document.

- ***Rake***: Rake stands for Rapid Automatic Keyword Extraction. It is an existing python implementation that uses the NLTK toolkit for the score calculations.

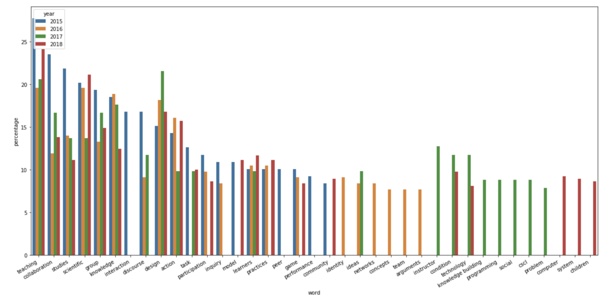

Figure 3: Trends of keywords over the years

How it works : * It splits the text into sentences and generates some words as candidates.
* Various punctuation signs will be treated as sentence boundaries. * All words listed in the stopwords file will be treated as phrase boundaries. This helps generate candidates that consist of one or more non-stopwords. However ,it won't work in cases where the stopword is part of the phrase. For example, 'new' is listed in RAKE's stopword list. This means that neither 'New York' nor 'New Zealand' can be ever a keyword. * For each keyword candidate generated, the algorithm computes the score of the pretended keyword, which is the sum of the scores for each of its words (if it's a composed word). The words are scored according to their frequency and the typical length of a candidate phrase in which they appear. * The last step is ranking the scores ansd selecting the ones with the highest score

Keywords in text: Select from the text the keywords already given by the authors of the paper.

Once we have all the keywords from the three methods, we try to select the most accurate keywords. So we create a method that selects the intersection of TF-IDF keywords and the Rake keywords with the keywords of the authors (if there are any keywords in the text), if the intersection is empty and we have the keywords in the text we only take the keywords of the text. If we don't have any keywords we take the intersection of the TF-IDF and the Rake. If again, the intersection is empty then we only take the TF-IDF keywords because they were closer to the text keywords than the Rake Keywords.

Once we have our final keywords, we create a method that goes into the metadata and assign to every document a publication year.

Since we don't have the same number of documents for each year we needed to scale over the years and use the percentage of appearence of a word. So after groupping the words by year and scaling them we found graph below 3

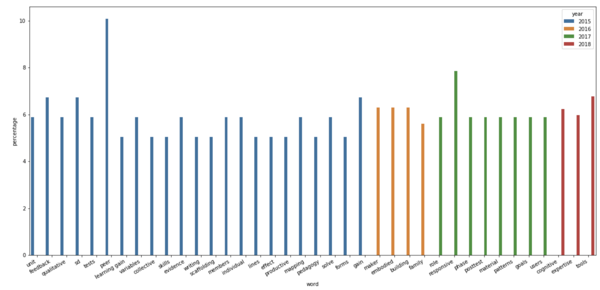

Figure 4: Distinct keywords over the years

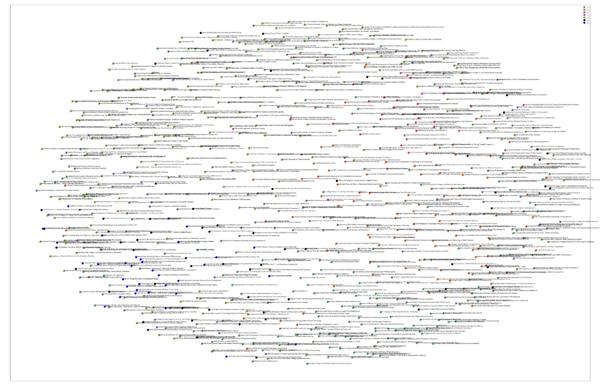

Figure 5: Baseline model document clusters

We wanted then to know what are the trends of words for each specific year, so we select distinct words for each year, it means that we select for each year the words that were present for that year and never for the others. We get the figure below. As we can see from the graph 4, the most used word in the papers, more than 10% of 2015 papers uses the word ‘peer’, more than 6% of the papers uses the words ‘building’, ‘embodied’ in 2016. In 2017 the most frequent keyword on is ‘reponsive’ and in 2018 ‘tools’ is present with more than 6%.

III. Methodology [find better section head..]

A. Document Clustering

To cluster the documents we decided to use two different approaches. The first one is Natural Langage Processing (NLP) and the second one is a network analysis. The most accurate one is the network clustering using the references graph.

1) Approach 1 : Natural Langage Processing (NLP):

a) Baseline : TF-IDF combined with K-means: With the natural langage processing, we start with a naive method which consists of clustering the documents using only TF-IDF. Each document has a vector representation with values for each word present in the document depending on the frequency of the word in this document regarding its frequency in all the other documents. So if a word is a keyword to the document and is not a common word than its value will be more important than the other words in the vector of the document. Once we have our vector representation, we select the max features parameter as 2500, it means that we will only consider the top max_features ordered by term frequency across the corpus. On this word representation we will run K-means.

K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k), resulting in the partitioning of the data into Voronoi cells. It can be considered a method of finding out which group a certain object really belongs to.

This method does not give good results as we can see it in the graph below 5 all the vectors are together there is no clear separation between the documents. So we decide to use other methods using word embeddings and properties of TF-IDF.

b) Second method : Average of GloVe vectors combined with K-means: The second method was to use pre-trained words embeddings. We decided to use ***GloVe*** library pre-trained on Wikipedia dataset.

GloVe, coined from Global Vectors, is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus (Wikipedia), and the resulting representations showcase interesting linear substructures of the word vector space. It is developed as an open-source project at Stanford.[8]

We construct a dictionnary from the words embeddings mapping each word of a document to its vector representation, than we average all the vectors of each document. Since each words has its representation in space, the average of all the words for each document will lead to a vector representing a document.

Once we get a document vector representation, we apply to it K-means. The results are better, but we have only two clusters. If we increase the number of clusters than there is no clear separation anymore separation. The plot below 6 reprensents the documents belonging to each cluster.

c) Third method : Weighted average of TF-IDF + GloVe combined with K-means: With this method, we did not use the parameter max features of the TF-IDF vector, we took all the vocabulary generated by all the documents and we created a method that sorts the document frequency vector, take the 500 first words and do a weighted average of the words embedding vectors with the TF-IDF score.

Once we get the weighted average vector, and since

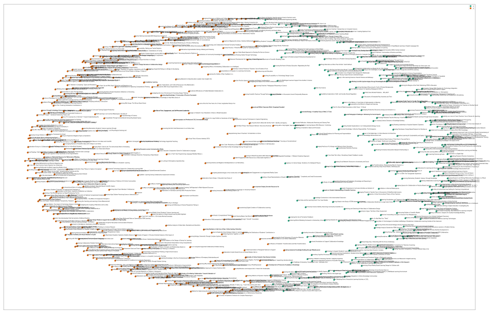

Figure 6: Cluster using Average of GloVe vectors combined with K-means

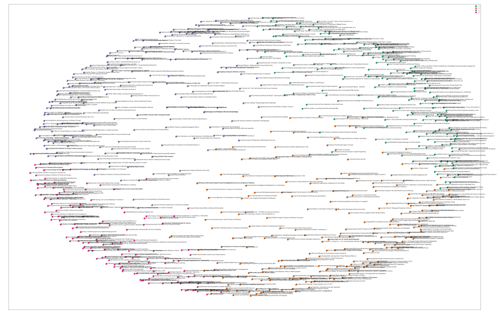

Figure 7: Cluster using Weighted average of TF-IDF + GloVe combined with K-means

we are using a words embedding of 50 dimentions. We decide to use feature reduction technique to reduce the number of dimensions and represent our vectors in an other vector space where all the features are orthogonal to each others.

To apply the feature reduction to our vectors we used ***Principal Component Analysis (PCA)*** algorithm.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. If there are ***n*** observations with ****p**** variables, then the number of distinct principal components is ***min(n-1,p)***. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables [9].

Once we get our new vector representation, we apply K-means and we get the best separation, with a clear separation of 4 clusters as we can see it in the figure below **??**.

2) Approach 2 : References Network Analysis: For the references graph we choose to construct 4 different graphs to be able to extract the more information we can. Our first purpose is to be able to cluster the documents that are about the same subject.

a) First reference graph : Intersection between referenced documents: Our first graph is constructed using the references of each document. It is constructed as follow, for each two documents, if there is an intersection between the references of both documents then there is a link. The weights of the edges are the length of the intersection set. The intuition behind is that the more the weight is high, the more the documents are likely to be about the same subject since they are citing the same documents. From this graph, we create a subgraph of it, we filter the edges keeping only those who have a weight strictly higher than 3. It means that, each node of the subgraph, is related to at least one other node with 4 same documents referenced.

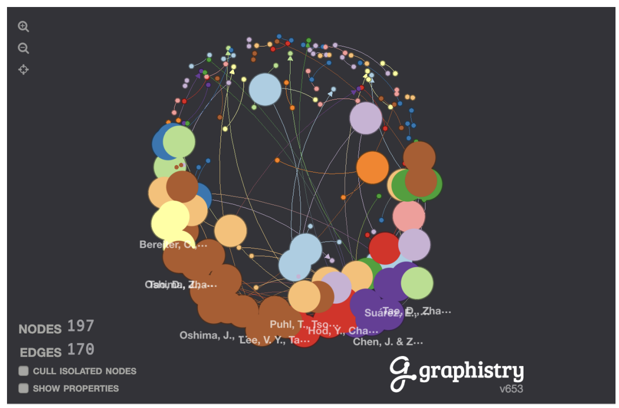

Figure 8: Intersection between references graph

The resulted graph is plotted in this figure 8, is an undirected graph with 805 nodes and 6699 edges. On this graph we extract two main information : * Strongly connected components to be able to detect clusters and related documents. * In_degree in order to check if there are documents that are related to a lot of documents and those documents would be more general than others, and won't necessarly belong to one specific cluster.

b) Second graph: Relations between references: The second graph is more about the referenced doc-

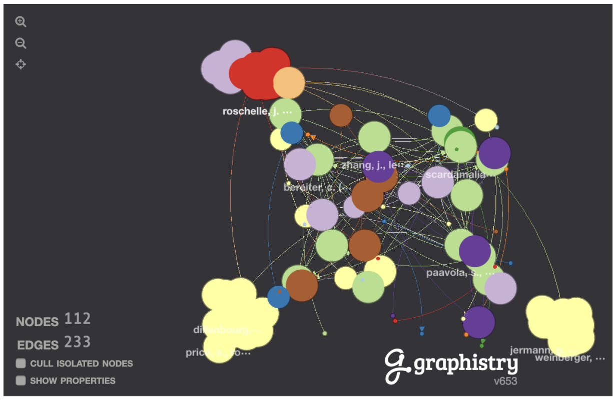

Figure 9

uments and not about the documents themselves. The aim of this graph is to see if there are documents that are more likely to cited together. If it is the case, then these documents are likely to be about the same subject. It is a weighted graph constructed as follow: for each document, we create a link between the references of the document. We create a dictionnary of links, if the link is already in the dictionnary then we increase the weight of the link. So for instance, if two references are present together in the 5 distinct documents, then the weight of the edge would be of 5. We obtain a graph of 10845 nodes and 148594 edges represented in this figure **??**

We can directly select from the interactive graph the nodes with degree higher than 100, these nodes represent the more cited documents, but we will construct a subgraph to specifically tackle these important referenced documents. On this graph, to be able to detect similar documents we also run Strongly connected components algorithm. On top of it, we inspected the documents with higher weights, we will explain more about what we extracted from this graph in the result section.

c) Third graph: exploratory references graph: With this graph we aim to find what are the most cited documents, what are the documents that are the most influencial and how citations evolve over time.

This is a directed graph constructed in this way. For each document, if the document cites another document than there is a link between the document and the cited one. We get a graph of 11650 nodes and 13243 egdes. We also construct a subgraph of this graph, which is a co-citations in conferences, but because a lack of data, we find only three documents that cites other documents.

On this graph, we run ***PageRank*** algorithm to be able to extract the most influential documents.

PageRank (PR)* is an algorithm used by Google Search to rank web pages in their search engine results. In our case, It measure PageRank is a way of measuring the importance of referenced papers. Compared to the in-degree analysis we get almost the same results since we have no out degree for almost every node.

B. *Co-authorship*

To get insight into collaborations within the community we build graphs based on co-authorship of papers. To build the graph structure and visualize it the python package *networkx* was used. When constructing the graph, we associate qualitative features –such as the name and institution of the author that is represented by a node– to the nodes and edges of the graph.

1) Graph Construction:: First, we built a co-authorship graph based on individual authors. Each node represents an author, and two nodes are connected if the authors collaborated on a paper. Other works [10], [11] have used a directed graph that only adds a link from the first author of the paper to all other authors. However, as we are more interested in collaboration patterns and less in finding the most influential first author, we build an undirected graph with edges between all co-authors. On the downside, this yields a very dense graph. Hence, we prune the initial graph by removing nodes that do not satisfy additional conditions. Intuitively it is clear that a singular collaboration on a paper does not necessarily indicate that two people even know each other. A first condition was to only link two nodes together if the authors corresponding to each node have appear as co-authors in more than one conference. The graph produced by this approach is much more sparse and contains much fewer authors. Additionally all connected components contain fewer than 50 authors, making it possible to consider each author individually. Such a component is shown in figure 10.

A second alternative approach was to consider the strength of bounds. In the graph authors are only linked if they have written at least 2 papers together. This method yields a much larger component –containing 441 authors– as this second condition is weaker. To get an understanding of author to author relations in this component we need to further prune away parts of the graph that are not insightful. When we look at individual author within the community, we are interested well collaborative individuals. Hence we remove nodes that have collaborated with fewer than 3 connections to other authors. The resulting graph can be seen in figure 11. Details creating such a plot can be found in the section dedicated to this below.

Focusing in on the authors is not the only way to interpret the co-authorship graph. One should also consider the overall structure of the graph. As graph is not connected, and contains one large main component with 1141 authors and 208 very small components with less

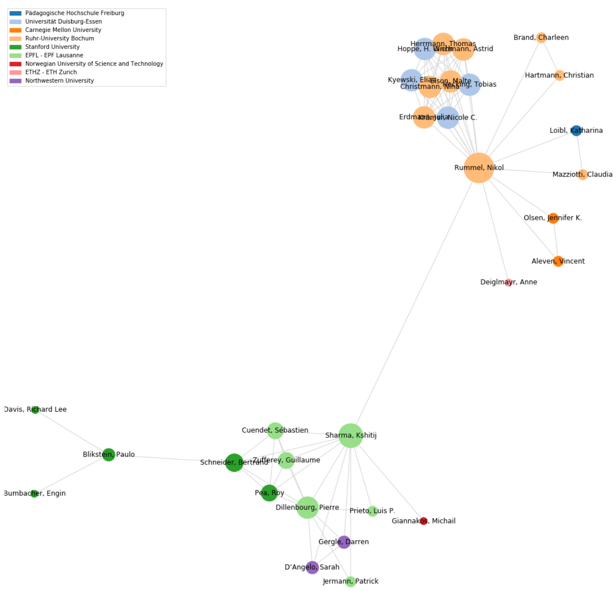

Figure 10: One of the connected components of the co-authorship graph

Figure 11: Largest Component

Figure 12: Subcommunities by Louvain

than 25 authors in them we focus on analyzing the structure of the main component of the graph. First we aim to detect sub-communities within the graph. Modularity is one measure used to find such communities. It measures the connection strength within a community relative to the strength of outgoing connections. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. A non-parametric algorithm for community detection is the Louvain Method [12]. We used the implementation of Thomas Aynaud [13]. This gives a good partitioning of the graph into 26 communities of median size 42.84. This is quite a large number of communities, so we consider an alternative way to yield bigger partitions. An alternate way to find subcommunities is to look at normalized cuts, which partition the nodes into groups with many within-group connections and relatively few between-group connections. The advantage of this method is that we can choose how many cuts to make, and we are thus able to define the size of subcommunities. Finding normalized cuts in a graph can be done using *sklearns* spectral clustering algorithm by passing it the adjacency matrix of the graph. With this method we are able to neatly separate the main component graph in three large communities. We note that the clusters found using method one are contained in the clusters found by method 2. See Figure 12 and Figure 13.

To further analyze these subcommunities we look at the distribution of nationalities and institutions over the years. Additionally, to understand differences in the

Figure 13: Subcommunities via normalized cuts

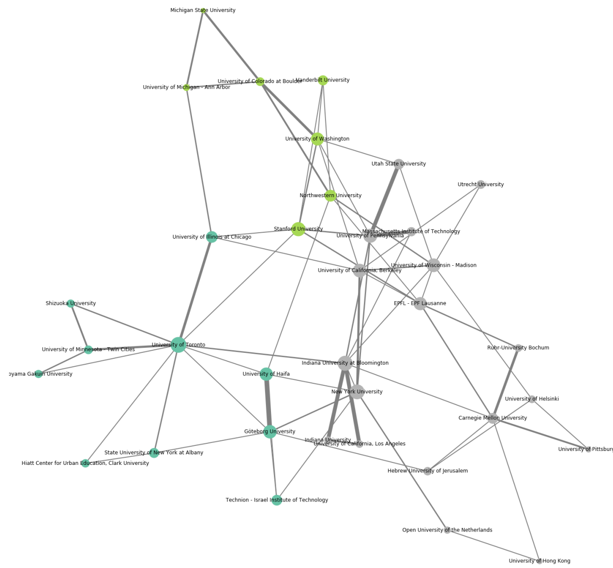

Figure 14: Different formats used to reference institution

research being conducted, we analyze the frequency of keywords used for the papers that were writing within that.

To get a better overview of how co-authorship happens across countries and institutions, we built graphs where every node represents a country respectively an institution. Two countries or institutions are linked if two authors from these two institutions collaborated on a paper together. The country graph is simple enough that it does not have to be pruned. The institution graph on the other hand is again not comprehensible without further processing. Thus, similar to the first approach on pruning presented, we only consider institutions who collaborated in at least two conferences. This the main component of this graph is shown in figure 14.

2) Numerical analysis of graphs:: To analyze the structure of the graphs themselves we calculated the following measures:

- Diameter: longest shortest path in the graph.
- Average clustering coefficient: Measure of how much the nodes in a graph tend to cluster together. Average over all clustering coefficients. The clustering coefficient of a node is the fraction of triangles prassing through the node over all possible triangles through that node [14].
- Density: the number of edges in the graph divided by the maximum number of possible edges for a graph with the same number of nodes

3) Plotting of co-authorship graphs:: Graphs were visualised with the built in plot functions and graph layouts of networkx. The force directed layout was used as it clusters together nodes with strong connections. Then the size of each node was changed based on the position of the node in the graph. Central nodes with a lot of connections should be larger, while nodes with fewer connections should be smaller. To achieve this we used the measures of degree centrality for graphs with few nodes and edges and betweenness centrality for larger graphs . Degree centrality measures the fraction of degrees (number of connections) a node has with respect to the total number of connections in the graph. Betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through it. Degree centrality is more adept for smaller graphs as their structure is still easy to interpret. As by construction betweenness takes into consideration the degrees of it neighboring nodes, it better structures the nodes large graphs into central and less central nodes.

C. *Creating Interactive Visualizations*

To better understand and interact with the data we create interactive graphs on top of the static graphs. Many libraries exist to allow for interaction. While working on the project, we worked with *bookeh*, *holoviews*, *mpld3* and *plotly*. For some time implementing a custom visualization using *D3* was also contemplated. However, functions provided by plotly is most general and allows to easily build interactive figures, while not being as heavy as booked and holoviews. Additionally, it is the most well maintained. D3 requires a running server to be viewed and has a hard time dealing with large graphs. To plot figures we create them in networkx to grab the

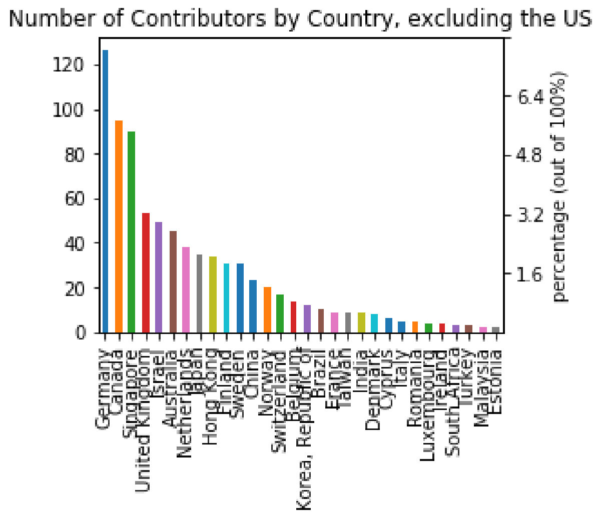

Figure 15: Different formats used to reference institution

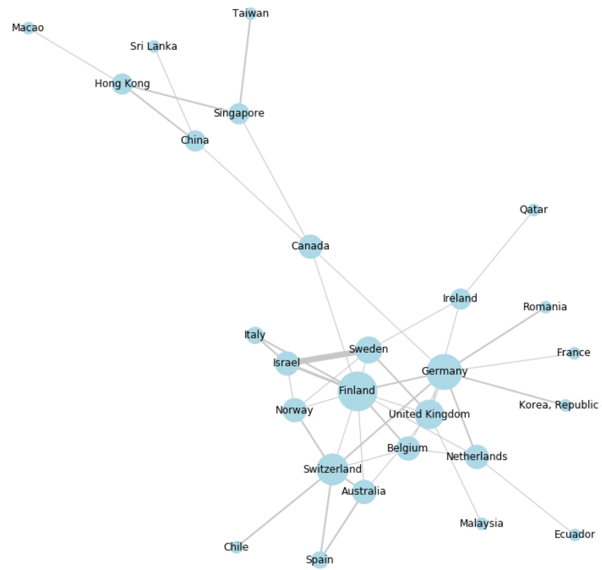

Figure 16: Collaborative papers between countries

Figure 17: Movement of participants across institutions

positioning of the nodes. Then we pass this positioning to plotly to plot.

IV. Results

In this section we present first insights gathered during the Bibliographic analysis.

A. Community Composition & Structure

We found that participants from 372 institutions participated over the last 4 years, with a median of 128 institutions per year. The number of institutions participating increased each year, with a noticeable peak in 2018, with participants from 261 institutions. This is a considerable increase from the last conference ICLS in 2016 which had 135 participating institutions. Most participants are from the US, with a average of 50% of all participants being from the US. We could also observe a trend of increase in US participants with only 35% of US participants in 2015 and 51% in 2017, followed by 64% in 2018. Overall, participants from 41 nations have participated over the 4 years, with an average of 28 nations represented at each conference. After the US, participants of Germany, Canada and Singapore make up the biggest proportion of participants **??**.

1) Collaboration: ***International Collaboration:** The U.S. again stands out as a major player, with most collaborations happening between the U.S. and other countries. Out of 851 papers, 153 were international collaborations. Out of these 153, 108 involved at least one U.S. participant. Excluding U.S. participation we find figure 23. We noted that Finland is very collaborative, and that there exists a strong bond between Sweden and Israel. Additionally we observed that Asian nations seem to collaborate almost exclusively with each other or Canada.

2) Movement of partipcants: Using the data we have of participants over the years, and knowing their affiliation, we can now find members that change institutions over the years. Over the years of recording we find 14 participants that have changed moved countries and 47 that have moved institutions. We can see that most movement occured within the US respectively Europe, 17.

B. Clustering and referecences graphs results

The distribution of referenced publications years are plotted in the graph 18. We can see from the graph that there is a kind of a trend, each two, three years there is a

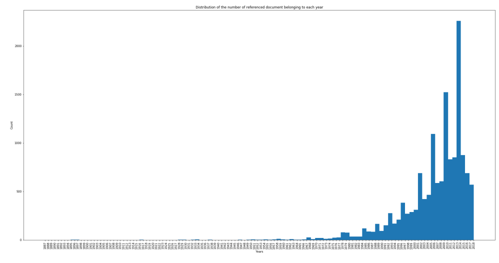

Figure 18: Distribution of the number of referenced documents belonging to each year

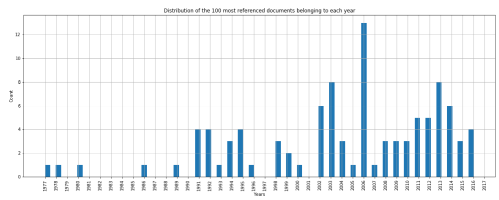

Figure 19: Distribution of the 100 most referenced documents belonging to each year

	Document	In_degree
0	jordan, b., & henderson, a. (1995). interaction analysis: foundations and practice	34
1	lave, j., & wenger, e. (1991). situated learning: legitimate peripheral participation	33
2	scardamalia, m. (2002). collective cognitive responsibility for the advancement of knowledge	28
3	ngss lead states. (2013). next generation science standards: for states, by states	28
4	national research council. (2012). a framework for k-12 science education: practices, crosscutting concepts, and core ideas	27
5	papert, s. (1980). mindstorms: children, computers, and powerful ideas	25
6	barron, b. (2003). when smart groups fail	22
7	scardamalia, m., & bereiter, c. (2006). knowledge building: theory, pedagogy, and technology	20
8	brown, j. s., collins, a., & duguid, p. (1989). situated cognition and the culture of learning	17
9	vygotsky, l. s. (1978). mind in society: the development of higher psychological processes	16

Figure 20: In-degree count for each document

	First_document	Second_document	Number_of_times_cited_together
0	bereiter, c. (2002). education and mind in the knowledge age	scardamalia, m. (2002). collective cognitive responsibility for the advancement of knowledge	22
1	scardamalia, m. (2002). collective cognitive responsibility for the advancement of knowledge	scardamalia, m., & bereiter, c. (2006). knowledge building: theory, pedagogy, and technology	19
2	barron, b. (2003). when smart groups fail	roschelle, j. (1992). learning by collaborating: convergent conceptual change	14
3	bereiter, c. (2002). education and mind in the knowledge age	scardamalia, m., & bereiter, c. (2006). knowledge building: theory, pedagogy, and technology	10
4	chi, m. t. h. (1997). quantifying qualitative analysis of verbal data: a practical guide	järvelä, s., & hadwin, a. f. (2013). new frontiers: regulating learning in cscl	9
5	bereiter, c. (2002). education and mind in the knowledge age	zhang, j., scardamalia, m., reeve, r., & messina, r. (2009). designs for collective cognitive responsibility in knowledge building communities	9
6	scardamalia, m. (2002). collective cognitive responsibility for the advancement of knowledge	zhang, j., scardamalia, m., reeve, r., & messina, r. (2009). designs for collective cognitive responsibility in knowledge building communities	9
7	sawyer, r. k. (2007). group genius: the creative power of collaboration	scardamalia, m., & bereiter, c. (2006). knowledge building: theory, pedagogy, and technology	7
8	collins, a., & halverson, r. (2009). rethinking education in the age of technology: the digital revolution and schooling in america	lave, j., & wenger, e. (1991). situated learning: legitimate peripheral participation	7
9	bodemer, d. (2011). tacit guidance for collaborative multimedia learning	janssen, j., & bodemer, d. (2013). coordinated computer-supported collaborative learning: awareness and awareness tools	7

Figure 21: Count of documents cited together

pick of documents referenced followed by a two or three years of stagnancy. We can also notice that throughouts the year each pick is higher than the older one, which can be explained by the fact that The International Society of the Learning Sciences community is investigating more and more tools to facilitate learning over the years.

Since the number of documents is increasing over the years, we also wanted to investigate more if the relevance of all the documents published is increasing too. Using Pagerank algorithm and the In-degree of each nodes to get the most influancial documents referenced of the citation graph that we constructed (graph3) before. We filtered the first 100 documents and then contructed a histogram of the counts. We obtain the figure 19:

We can see that in the top 100 cited documents more than 12% are from 2006, 8% are from 2013, 6% from 2014… We can not say that we have the same trend as before, and the relevance of the documents is increasing in the same way than with the number of publications but what we can say is that more than 50% of the most cited documents are from 2006 until 2016, we is 10 years. The other half is between 1977 until 2005 which represents more than 25 years. On top of that, 2016 is the most important year in term of publications and maybe some documents did not have enough time to be cited yet.

The top 10 of the most influencial and most important referenced papers are represented in table 20:

We also noticed from the first graph that some documents were more likely to cited together. We can say that the more two documents are likely to cited together the more it is probable that both documents are about the same subject From the table 21 we can also notice that some papers cites like to cite the same authors together, which is logic in a sense, because if some authors always treat about the same subjects then they are experts in that field and a lot of their publications are relevant and can be cited together.

Using this graph, we also want to extract if there are some clusters of papers cited together. As we explained before, we run strongly connected algorithm to check if some cited documents are more likely to be cited together, and this will give us clusters of cited documents that are more likely to be about the same subjects. We can visualise one cluster, the most important one.

The largest strongly connected component is the composed of 8 nodes and 56 edges. This documents are all cited together more than 4 times. You can investingate the edges and the weights in the interactive graph constructed above. We wanted all the documents to be linked to each others to be more precise, but if we only look to giant connected component, we will get a network of 27 referenced paper with 142 egdes.

In the table 25 below are the documents of the giant connected component represented in the interactive graph 26.

Now we want to apply the same logic with the papers in-conferences. So we will focus on the analysis of the first graph that uses the intersection between the

		Documents
cluster 4	0	schommer, m. (1990). effects of beliefs about the nature of knowledge on comprehension
	1	stathopoulou, c., & vosniadou, s. (2007). exploring the relationship between physics-related epistemological beliefs and physics understanding
	2	hofer, b. k., & pintrich, p. r. (1997). the development of epistemological theories: beliefs about knowledge and knowing and their relation to learning
	3	chinn, c. a., buckland, l. a., & samarapungavan, a. l. a. (2011). expanding the dimensions of epistemic cognition: arguments from philosophy and psychology
cluster 3	0	yoon, s., klopfer, e., wang, j., sheldon, j., wendel, d., schoenfeld, i., scheintaub, h., & reider, d. (2013). designing to improve biology understanding through complex systems in high school classrooms: no simple matter! in the proceedings of the computer supported collaborative learning, madison, wisconsin
	1	ertmer, p.a., ottenbreit-leftwich, a.t., sendurur, e., & sendurur, p. (2012). teacher beliefs and technology integration practices: a critical relationship
	2	epstein, j. (2008). 'why model
	3	national research council. (2012). education for life and work: developing transferable knowledge and skills in the 21st century
	4	thompson, j, windschitl, m., & braaten, m. (2013). developing a theory of ambitious early-career teacher practice
	5	desimone, l. (2009). improving impact studies of teachers' professional development: toward better conceptualizations and measures
	6	yoon, s., koehler-yom, j., anderson, e., lin, j., & klopfer, e. (accepted). using an adaptive expertise lens to understand the quality of teachers' classroom implementation of computer-supported complex systems curricula in high school science
	7	mueller, j., wood, e., willoughby, t., ross, c., & specht, j. (2008). identifying discriminating variables between teachers who fully integrate computers and teachers with limited integration
cluster 2	0	dougiamas, m., & taylor, p. (2003). moodle: using learning communities to create an open source course management system
	1	coopey, e., schneider, l., & danahy, e. (2013). interlace: interactive learning and collaboration environment
	2	scardamalia, m., & bereiter, c. (2003). knowledge building environments: extending the limits of the possible in education and knowledge work
	3	scardamalia, m., & bereiter, c. (1996). student communities for the advancement of knowledge
cluster 1	0	brown, j. s., collins, a., & duguid, p. (1989). situated cognition and the culture of learning
	1	engeström, y. (1987). learning by expanding
	2	lave, j., & wenger, e. (1991). situated learning: legitimate peripheral participation: cambridge univ pr
	3	vygotsky, l. s. (1978). mind in society (m
	4	cole, m. (1996). cultural psychology: a once and future discipline: harvard university press
	5	shaffer, d. w., hatfield, d., svarovsky, g. n., nash, p., nulty, a., bagley, e., . . . mislevy, r. (2009). epistemic network analysis: a prototype for 21st-century assessment of learning

Figure 22: Table of the documents belonging to the same clusters

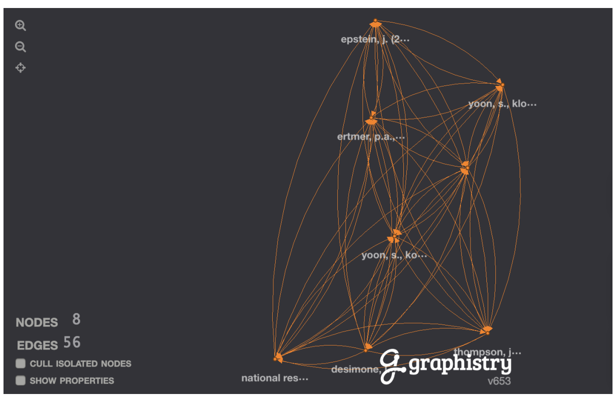

Figure 23

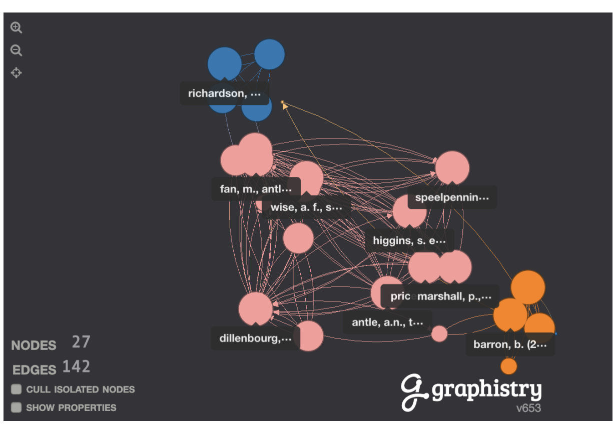

Figure 24: Giant component graph

	Documents
0	white, t., & pea, r. (2011). distributed by design: on the promises and pitfalls of collaborative learning with multiple representations
1	schneider b., wallace j., pea, r. & blikstein p. (2013). preparing for future learning with a tangible user interface: the case of neuroscience
2	dillenbourg, p., & evans, m. (2011). interactive tabletops in education
3	speelpenning, t., antle, a. n., doering, t., & van den hoven, e. (2011). exploring how tangible tools enable collaboration in a multi-touch tabletop game
4	price, s., rogers, y., stanton, d., & smith, h. (2003). a new conceptual framework for cscl
5	fan, m., antle, a. n., neustaedter, c. & wise, a. f. (2014). exploring how a co-dependent tangible tool design supports collaboration in a tabletop activity
6	jordan, b., & henderson, a. (1995). interaction analysis: foundations and practice
7	rogoff, b. (1990). apprenticeship in thinking: cognitive development in social context
8	kirschner, p. a., sweller, j., & clark, r. e. (2006). why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching
9	wise, a. f., saghafian, m. & padmanabhan, p. (2012). towards more precise design guidance: specifying and testing the functions of assigned student roles in online discussions
10	roschelle, j. (1992). learning by collaborating: convergent conceptual change
11	tomasello, m. (1995). joint attention as social cognition
12	jermann, p., mullins, d., nüssli, m.-a., & dillenbourg, p. (2011). collaborative gaze footprints: correlates of interaction quality
13	higgins, s. e., mercier, e., burd, e., & hatch, a. (2011). multi-touch tables and the relationship with collaborative classroom pedagogies: a synthetic review
14	schneider, b., & pea, r. (2013). real-time mutual gaze perception enhances collaborative learning and collaboration quality
15	meier, a., spada, h., & rummel, n. (2007). a rating scheme for assessing the quality of computer-supported collaboration processes
16	barron, b. (2003). when smart groups fail
17	antle, a.n., tanenbaum, j., macaranas, a. and robinson, j. (2014) games for change: looking at models of persuasion through the lens of design
18	ullmer, b. & ishii, h. (2000). emerging frameworks for tangible user interfaces
19	shaer, o., strait, m., valdes, c., feng, t., lintz, m., & wang, h. (2011). enhancing genomic learning through tabletop interaction
20	fernaeus, y. & tholander, j. (2006). finding design qualities in a tangible programming space
21	richardson, d. c., & dale, r. (2005). looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension
22	lave, j. & wenger, e. (1991). situated learning: legitimate peripheral participation
23	schneider, b., jermann, p., zufferey, g., & dillenbourg, p. (2011). benefits of a tangible interface for collaborative learning and interaction

Figure 25: Documents belonging to the giant component

		Documents
cluster 9	0	Mazziotti, C., Rummel, N., & Deiglmayr, A. (2016). Comparing Students' Solutions When Learning Collaboratively or Individually Within Productive Failure
	1	Hartmann, C., Angersbach, J., & Rummel, N. (2015). Social Interaction, Constructivism and their Application within (CS)CL Theories In Lindwall, O
	2	Mazziotti, C., Loibl, K., & Rummel, N. (2015). Collaborative or Individual Learning within Productive Failure: Does the Social Form of Learning Make a Difference
	3	Hartmann, C., Rummel, N., & Loibl, K. (2016). Communication Patterns and Their Role for Conceptual Knowledge Acquisition From Productive Failure
cluster 8	0	McBride, E., Vitale, J., & Linn, M. (2018). Learning Design Through Science vs
	1	McBride, E., Vitale, J., Applebaum, L., & Linn, M. (2017). Examining the Flow of Ideas During Critique Activities in a Design Project In Smith, B
	2	McBride, E. A., Vitale, J. M., Applebaum, L., & C.Linn, M. (2016). Use of Interactive Computer Models to Promote Integration of Science Concepts Through the Engineering Design Process
	3	McBride, E., Vitale, J., & Linn, M. (2018). Middle School Student Ideas on the Relative Affordances of Physical and Virtual Models
	4	Mishra, S. & Iyer, S. (2018). Promoting Cognitive Processes of Knowledge Integration
cluster 7	0	Greenhow, C., Hershkovitz, A., Baruch, A. F., Askari, E., Tsovaltzi, D., Asterhan, C., Puhl, T., Weinberger, A., Bouton, E., & Polman, J. (2016). Teachers and Professional Development: New Contexts, Modes, and Concerns in the Age of Social Media
	1	Erkens, M., Schlottbom, P., & Bodemer, D. (2016). Qualitative and Quantitative Information in Cognitive Group Awareness Tools: Impact on Collaborative Learning
	2	Schnaubert, L. & Bodemer, D. (2016). How Socio-Cognitive Information Affects Individual Study Decisions
	3	Asterhan, C. & Bouton, E. (2017). Secondary school peer-to-peer knowledge sharing through social network technologies In Smith, B
	4	Erkens, M. & Bodemer, D. (2017). Which Visualization Guides Learners Best
	5	Tsovaltzi, D., Dutta, N., Puhl, T., & Weinberger, A. (2017). Group and Individual Level Effects of Supporting Socio-Cognitive Conflict Awareness and Its Resolution in Large SNS Discussion Groups: A Social Network Analysis In Smith, B
	6	Heimbuch, S. & Bodemer, D. (2016). Effects of Implicit Guidance on Contribution Quality in a Wiki-Based Learning Environment
	7	Puhl, T., Tsovaltzi, D., & Weinberger, A. (2015). A Long-Term View on Learning to Argue in Facebook: The Effects of Group Awareness Tools and Argumentation Scripts In Lindwall, O
cluster 6	0	Dornfeld, C. L. & Puntambekar, S. (2016). Negotiation Towards Intersubjectivity and Impacts on Conceptual Outcomes
	1	Evenstone, A. L. & Puntambekar, S. (2015). Internalization of Physics Concepts and Relationships Based on Teacher Modeling of Collaborative Prompts In Lindwall, O
	2	Dornfeld, C., Zhao, N., & Puntambekar, S. (2017). A Mixed-Methods Approach for Studying Collaborative Learning Processes at Individual and Group Levels In Smith, B
	3	Martin, N. D., Gnesdilow, D., & Puntambekar, S. (2015). Peer Scaffolding to Learn Science in Symmetrical Groups Collaborating Over Time In Lindwall, O

Figure 26: Documents in conferences belonging to the same clusters

referenced documents of each two by two papers. The length of the intersection set had to be more than 4.

So if more than 4 documents are referenced by two distinct documents than the latter are more likely to be about the same subect. In the table 26 we can find some of the papers that belong to the same clusters.

When applying this kind of clustering we find that most of the clusters contains almost the same authors with different papers, which is logic because if they are specialists in one topic and write several papers on it, they might reuse some references in several documents.

V. Conclusions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

REFERENCES

[1] Isls website. https://www.isls.org/conferences.
[2] poppler. https://poppler.freedesktop.org/.
[3] Isls author guidelines. https://cscl2019.com/upload/ISLS_Author_Guidelines.pdf, 2014.
[4] Scholarcy reference extraction api. https://ref.scholarcy.com/api/.
[5] Apa formatting and style guide by purdue writing lab. https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/reference_list_articles_in_periodicals.html.
[6] University domains and names data list and api. https://github.com/Hipo/university-domains-list, 2018.
[7] Tf-idf definition. https://en.wikipedia.org/wiki/Tf%E2%80%93idf.
[8] Glove definition. https://en.wikipedia.org/wiki/GloVe_(machine_learning).
[9] Pca definition. https://en.wikipedia.org/wiki/Principal_component_analysis.
[10] France Cheong and Brian J Corbitt. A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. *PACIS 2009 Proceedings*, page 23, 2009.
[11] James W Hesford, Sung-Han Sam Lee, Wim A Van der Stede, and S Mark Young. Management accounting: a bibliographic study. *Handbooks of management accounting research*, 1:3–26, 2006.
[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
[13] Thomas Aynaud. Louvain community detection. https://github.com/taynaud/python-louvain, 2018.
[14] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.