What Are We Talking About?

An Analysis of the SIGCSE-Members Listserv

Austin Cory Bart
Virginia Tech
Blacksburg, VA
acbart@vt.edu

ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam sagittis arcu sed mollis commodo. Donec et augue pretium, tincidunt neque sed, vehicula ex. Vestibulum quis commodo libero. Suspendisse potenti. Nulla varius turpis enim, non dapibus leo commodo non. Maecenas fringilla dictum purus nec egestas. Sed vulputate mauris at mauris lobortis mattis. Suspendisse consequat est vitae turpis commodo, at dignissim turpis dapibus. Vivamus vel volutpat orci, at tempor nibh. Praesent in tempor leo. Pellentesque pulvinar venenatis nisl. Duis congue libero vel tellus interdum, vel luctus justo iaculis. Sed sagittis elit ut arcu sodales posuere. Proin vitae tellus eleifend, varius magna vel, porta neque. Phasellus hendrerit, sapien eget egestas tempor, risus nunc congue nibh, in congue enim ex bibendum odio.

Donec placerat ipsum a imperdiet sagittis. Etiam malesuada rutrum sem vel faucibus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Curabitur nec lacus tempus, tincidunt diam ut, efficitur sem. Duis luctus bibendum iaculis. Mauris iaculis fringilla fermentum. Donec consectetur sagittis neque in tincidunt. Maecenas vitae bibendum nisl. Curabitur sapien leo, consequat vel enim quis, porta auctor tortor. Etiam sed condimentum nisl, quis eleifend tellus. Suspendisse eu diam molestie, viverra nisi non, pretium dolor. Nunc sed nulla vel augue pellentesque cursus.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference format:
Austin Cory Bart. 1997. What Are We Talking About?. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97),* 5 pages.
https://doi.org/10.475/123_4

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

About 21 years ago, the SIGCSE-Members listserv began archiving posts by the Computer Science Education community.

This paper characterizes the history and state of the

To my knowledge, this is the first attempt at an comprehensive, quantitative analysis of the mailing list.

Kim Bruce published a summary of a particularly fervent conversation from the mailing list in an ITiCSE working group [1].

1.1 Audience

The primary audience for this paper are newcomers to the SIGCSE-Members listserv who would benefit from an introduction. This particularly includes new researchers and teachers who are

However, a number of our contributions should be of interest to existing members who want a more quantitative insight into the community.

1.2 Contributions

This paper makes a number of contributions.

- Characterization of trends over time in posting behavior,
- An analysis of job posting behavior,
- An examination of the word usage, and
- Modeling thread behavior and popularity.

2 DATA COLLECTION

The SIGCSE-Members listserv is restricted to posting by confirmed members, but maintains a public archive of all posts. This public archive was scraped, processed, and analyzed using a combination of Python scripts, which are available at. The following data was collected from the archive:

- The body of the email (either as unicode text or HTML).
- The subject line of the email.
- Any attachments, including their MIME filetype.
- The name of the sender (but not their email, as described below).
- The timestamp of the post.
- The conversation thread that the email belongs to (and its position within the thread).

The archive does not reveal posters' emails unless the viewer logs in as a confirmed member. In the spirit of privacy, we opted to only collect data that was publicly viewable without logging in. This means that, although we collected user-friendly names of users, we were not able to collect the exact email address that sent each email. Section 6.1 describes the methodology used to overcome this limitation.

Once the data was downloaded, it was processed. This includes the conversion of special unicode characters to simplify textual analysis. HTML files were converted to a plain-text representation using the Html2Text library

2.1 Descriptive Statistics

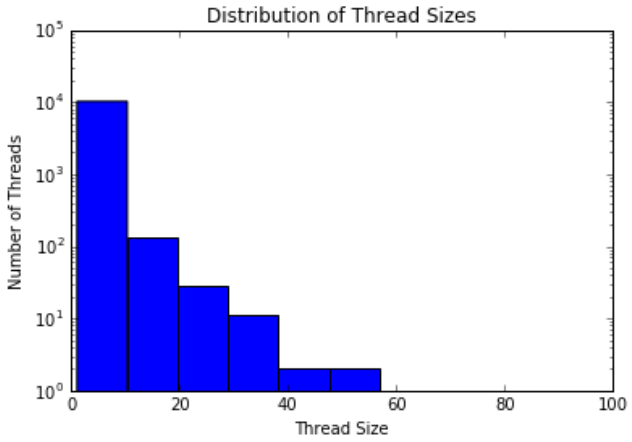

Figure 1: Distribution of Thread Sizes

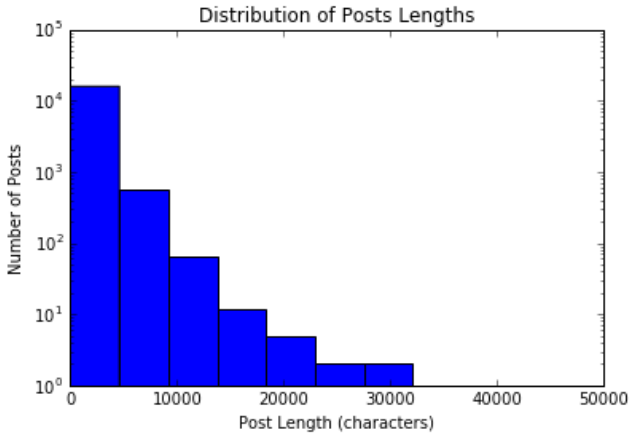

Figure 2: Distribution of Post Lengths

Figure 1 and 2 shows the distribution of email thread sizes and individual email length, respectively. Note that both graphs have logarithmic scales.

Most threads tend to be very small, lasting only one or two posts.

Similarly, most emails tend to be very short, lasting less than 5000 characters.

2.2 Thread Duration

Remarkably, the duration of non-trivial threads follows a power law almost exactly, specifically

$$\%\ of\ Threads\ Ended = \frac{1}{2.1^{days}}$$

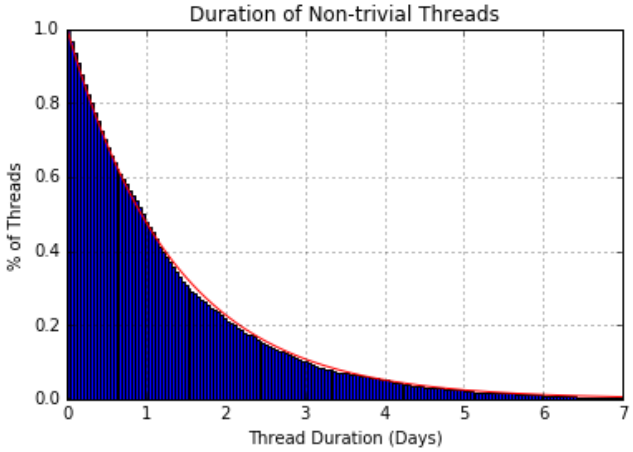

Figure 3: Non-trivial Thread Duration

In fact, the Mean Squared Error is very close to 0.

The implication is fairly straightforward: almost half (47.6%) of all non-trivial threads end after a day, roughly a quarter (22.7%) end after two days, a tenth (10.8%) after three days, and so on.

Previous studies of forum and email thread lengths suggest that power-law distributions are common in these kinds of conversations.

3 TIME

Figure 4 shows trends over time in user posting behavior. Each of the four graphs show a different granularity of time: over the entire time period, grouped by month, grouped by day of the week, and grouped by the time of day.

The Yearly trend shows a chaotic growth over the past two decades. The variation over time is considerably higher than in other graphs. However, linear regression reveals a significant ($p < .01$) but tiny ($slope = .13$) positive trend. This suggests that the rate of posting has relatively increased over the years, but with strong fluctuation – most likely due to the cyclical boom/bust cycles seen in computer science programs. The most sustained activity seems to have occurred during the 2009-2010 time period, while the lowest was during the 2001-2003 time period.

The Monthly distribution reveals that users are disengaged during the summer months and during the winter holidays. Although January quickly peaks with new posts, the rest of the spring is a slow downward trend until posts fall off completely in June. When most US Fall semesters begin in September, posts also pick-up and eventually peak in October, before falling back down in November. It is unclear exactly why posts peak so heavily in January, but two hypotheses are: 1) anticipation of the SIGCSE conference, or 2) to make up for lost time during December. Of course, there is high variation within most months, so these conclusions are tenuous.

The day of week graph shows the frequency of different numbers of emails sent during days of the week. This graph indicates that for any given day of the week, there are usually no posts. When posts do occur, they tend to happen considerably more often on weekdays than on weekends. This suggests that most posters disengage during

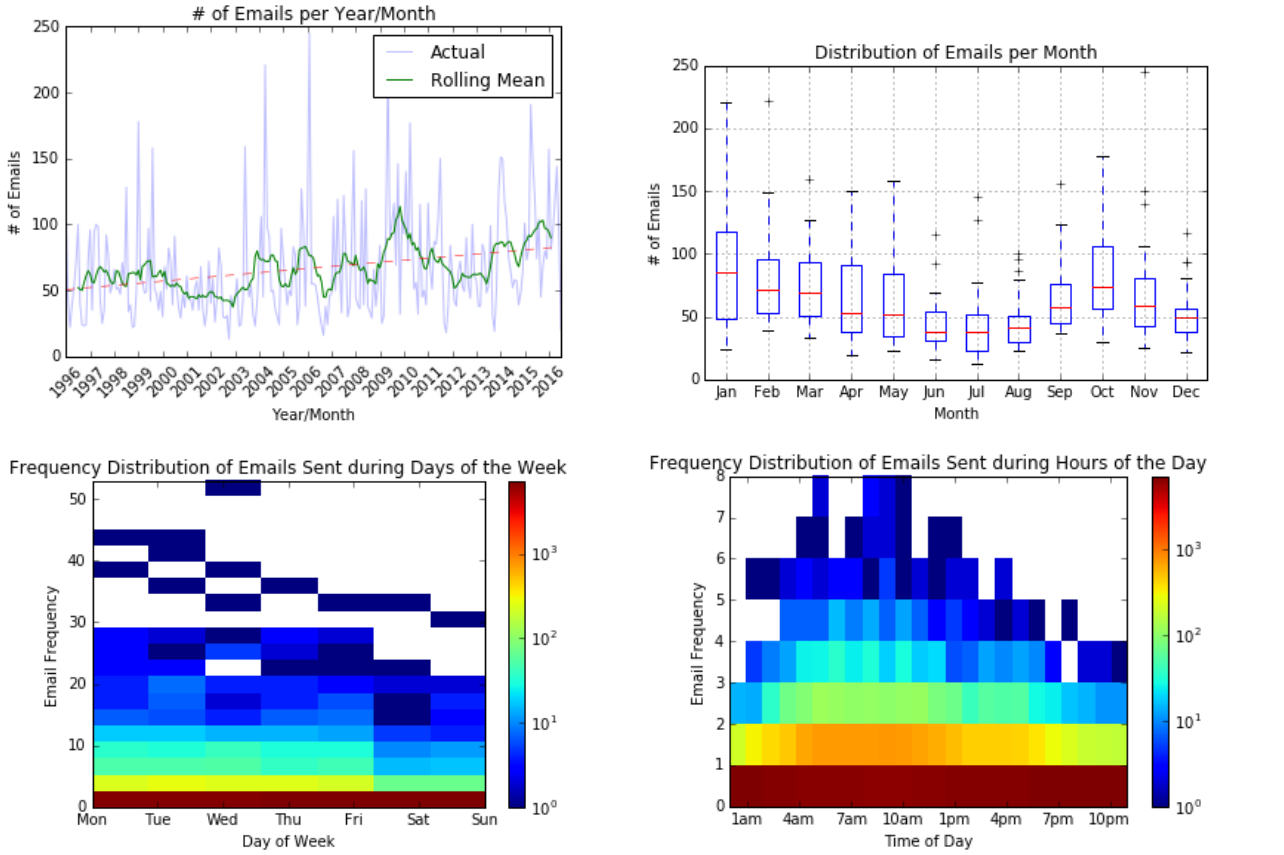

Figure 4: Posts over Time

the weekend. There appears to be a generally negative trend during the week, at least in terms of the heavily active days. If accurate, then this implies that posters steadily disengage over the course of the week.

The final graph shows the frequency of posts over hours of the day. Similar to the day of week graph, most hours have no emails posted. However, the graph does seem to show three general time periods: a busy morning period, a moderately busy evening period, and a quiet night period. One conclusion is that SIGCSE posters tend to catch up on their correspondence early in the day and trail off over time, mostly disengaging after work. Timezone data was not always available for each post. When it was available, all times were adjusted to Greenwich Mean Time in order to meaningfully compare post times from a common vantage point. Therefore, the 24-hour graph should be taken as an approximation of the actual hourly posting behavior.

These posting habits reveal human behavior within the SIGCSE-members listserv. Although not particularly surprising, participants tend to avoid posting during non-work hours, on weekends, and during holiday seasons. These patterns should be considered when waiting on responses or when trying to decide when to post messages.

4 WORD USAGE

We analyzed the usage of certain words over time. First, we looked at programming language names. Then, we looked at gendered vocabulary. Both analyses were performed at the year level, due to the relatively limited number of occurrences.

4.1 Trending Programming Languages

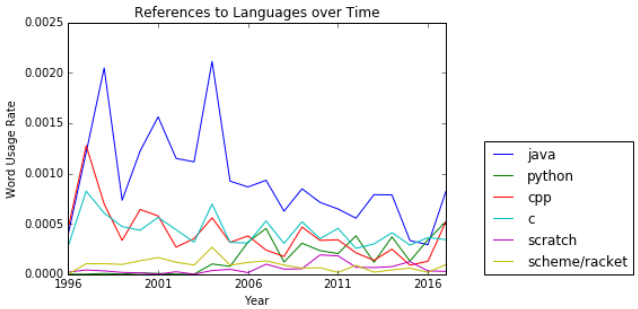

Figure 5: Programming Language Usage Rate over Time

Java has almost always been the top language C++ matched it the first couple years, but it has since dwindled C has consistently performed near the top Python has grown dramatically in popularity Surprisingly little conversation about Scratch or Scheme/Racket

4.2 Gendered Language

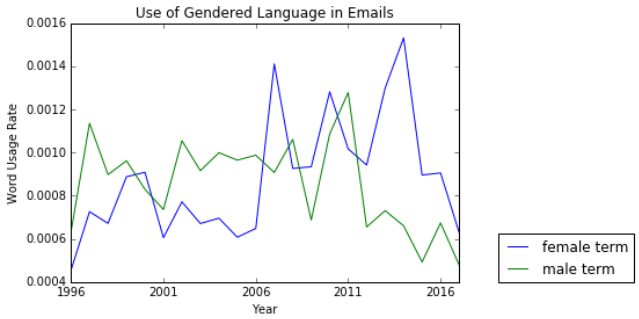

Figure 6: Gendered Word Usage Rate over Time

Figure 6 shows the trends in the use of gendered language over years. Two kinds of words were counted as gendered: "Male" and "Female". Male words include "men", "boys", "gentlemen", "his", "he"; female words include "women", "girls", "ladies", "her", and "she". The number of occurrences was then divided by the total number of words from that year in order to provide a time-adjusted rate.

In the first decade of the mailing list, most years had more male words used than female. However, that trend changed around 2007, which saw the beginning of a more commingled period. In the past half-decade, there has been considerably more female gendered words used than male. I suggest this trend reflects the increased emphasis within the SIGCSE community on the gender balance problem.

5 JOBS AND CONFERENCES

Machine Learning was used to classify each post based on the type of post. 300 posts were tagged as either "Normal Conversation", "Job Ad", or "Conference Discussion". These posts were then used as training data for a Support-Vector Machine using the body of the post as a Bag-of-Words. Metrics show the classifier to be highly effective at identifying different types of emails.

5.1 Trends

Figure 7 shows the yearly trend in job, conference, and regular posts.

Job advertisements have increased over the past decade.

The low points of the job advertisements actually correspond closely to the highest peaks of unemployment within the US economy.

Figure 8 shows the monthly trend in job, conference, and regular posts.

Job ads appear to be somewhat seasonal.

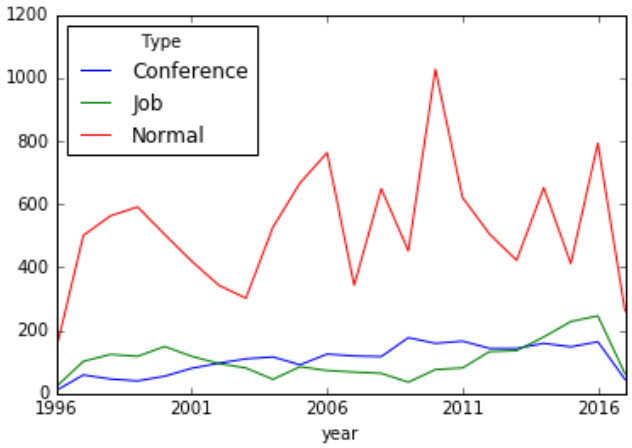

Figure 7: Post Types by year

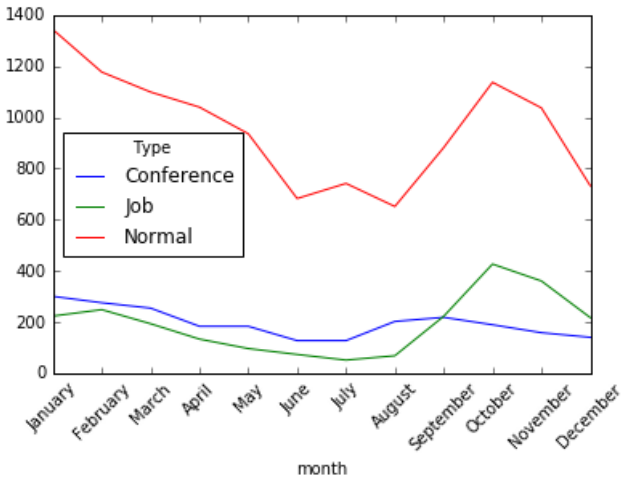

Figure 8: Post Types by month

6 POSTERS

6.1 Sender Name Normalization

Because specific email addresses were not available, it became necessary to parse the names associated with the posts. Unfortunately, these names are highly irregular. When users posted using different email addresses, or even if they simply changed their name within the system, a new name would be created.

To overcome this limitation, we normalized sender names. First, names were converted to lower case. Second, unnecessary titles were stripped out (e.g., "dr" or "doctor"). Third, each full name was split into a set of names and sorted (e.g., the name "charles babbage" would be converted to the list "babbage, charles") in order to avoid common reorderings of names.

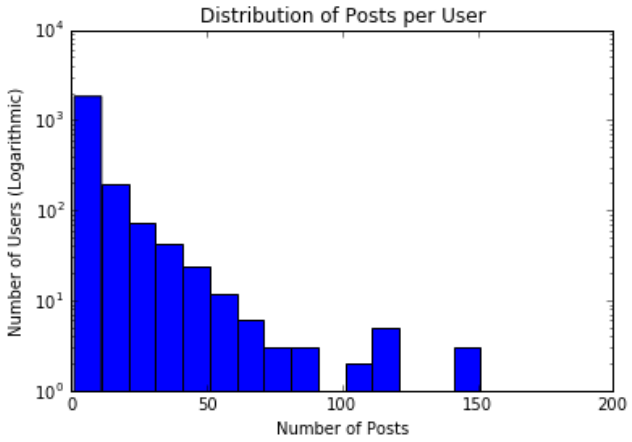

Figure 9: Distribution of Posts per User

6.2 User Activity

Figure 9 shows that most users had very few posts. In fact, the median number of posts per user is 2.

Top poster had approximately 2% of all the posts.

Any user that made more than 20 posts was considered "active". Any user that had more than X posts was considered a "super" user.

# of posts: 16525
# of users: 2205
# of supers 40
Supers as % of users 1.8140589569160999
Supers' % of posts: 25.706505295007563
# of actives 177
Actives as % of users 8.07
Actives' % of posts: 49.6

6.3 Activity and Productivity

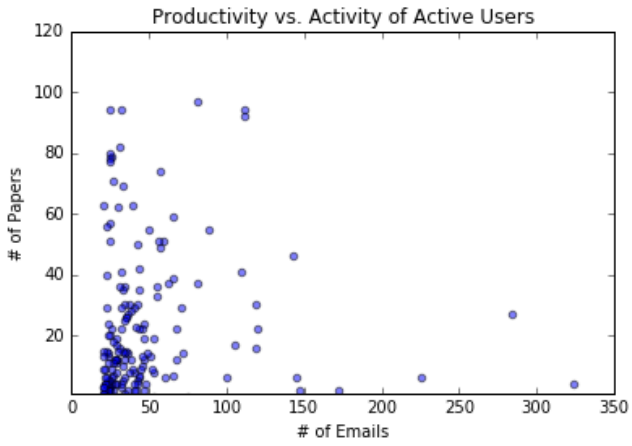

Figure 10: Productivity vs. Activity of Active Users

Each active member of the mailing list was identified and researched in order to correlate their mailing list activity with their productivity. The number of papers published within the Special Interest Group on Computer Science Education was used as a measure of productivity within the community. Note that this includes several publications, including SIGCSE, ICER, ITiCSE, and Koli. Figure 10 is a scatter plot showing the relationship between the users' posts and their SIGCSE publications. Surprisingly, there was almost no correlation between activity and productivity. There were a number of users who published frequently without posting to the mailing list, and many active users with few publications to their name at SIGCSE.

7 THREAD POPULARITY REGRESSION

An attempt was made to characterize what makes a popular thread. What makes a thread more popular? Does posting time/length/questions make a difference?

First, a stepwise linear regression model was attempted.

Second, each continuous variable was classified into finite categories and used in a logistic regression model.

However, in both cases, no models could be found to reasonably fit the data. In fact, the R^2 ends up being $< .01$ in all such models.

This suggests that, despite patterns in user behavior, there are no predictable ways to increase user response.

8 FUTURE WORK

8.1 Detailed User Analysis

As described in Section 6.1, exact email addresses were not available in the public archive, which hampered our ability to conduct fine-grained analysis of users. A further analysis of this archive would benefit from using the associated email data to more precisely identify users.

9 CONCLUSIONS

We have characterized the SIGCSE Members mailing list in a number of ways. We hope that this reveals some deeper insight.

REFERENCES

[1] Kim B Bruce. 2004. Controversy on how to teach CS 1: a discussion on the SIGCSE-members mailing list. In *ACM SIGCSE Bulletin*, Vol. 36. ACM, 29–34.