where $Len(S)$ denotes the length of a sentence S and is simply computed by counting the number of words appearing in it. Now, assume that the papers $Q_1, Q_2, \ldots, Q_n$ are citing paper P in the t^{th} year after publication of P. We define the average citeWords metric for paper P for the t^{th} year as:

$$Average\ citeWords(P,t) = \frac{\sum_{j=1}^{n} citeWords(P,Q_i)}{n} \quad (4)$$

Using Table 2, average citeWords value for paper *P* for the third year after publication (year 2008) can be calculated as:

$(citeWords(P, 6413388) + citeWords(P, 5052733))/2$

To compute $citeWords(P, 5052733)$, we see that paper 5052733 cites P in one citation context when a total of two papers are cited. Thus

$citeWords(P, 5052733) = \frac{11}{2} = 5.5$, where 11 is the length of the citation context.

Similarly, paper 6413388 cites paper P twice but in both the citation contexts, two papers are cited. Therefore,

$$citeWords(P, 6413388) = \frac{25}{2} + \frac{16}{2} = 20.5$$

Thus, $Average\ citeWords(P, 3) = \frac{20.5+5.5}{2} = 13$.

4.3 Correlation between citation counts and citation content features over the years

We investigate whether the average countX and average citeWords values over the years are correlated with the number of citations a paper receives. We reiterate that both average countX and average citeWords are normalized with respect to the number of citations received by the paper. We divide the set of papers in our dataset into 6 buckets based on the following criterion on the number of citations.

Bucket 1: Top 0.1% papers – citations 389-7859

Bucket 2: Top 0.1 - 1% papers – citations 95-389

Bucket 3: Top 1 - 5% papers – citations 29-95

Bucket 4: Top 5 - 10% papers – citations 16-29

Bucket 5: Top 10 - 25% papers – citations 6-16

Bucket 6: Rest of the papers – citations 0-6

For each of the Citation buckets, we plot the temporal profile for the average countX values, averaged for all the papers within that bucket, in Figure 2. The $X-$axis denotes the year after publication for the paper, ranging from 0 (same year as publication) to 10 (10^{th} year after publication). While averaging for a citation bucket for a particular year, we consider only those papers which have non-zero citations in that year. Minimum value of countX can be 1 for any citation edge. Interestingly, as per our hypothesis, various citation ranges show differences in terms of the average countX values. Some important observations from Figure 2 are:

1. There is an increase in value of countX in initial years irrespective of the citation bucket, and it further decreases continuously over the years. A slight increase is observed for the 10^{th} year after publication.

2. Highly cited papers are cited more number of times in a single paper.

We clearly see a correlation between the number of citations and the average countX profiles of the papers. Further, we investigate whether the countX values can discriminate between the 6 citation categories identified in [7]. Accordingly, we divided the set of papers into 6 categories mentioned in [7]. For readability, the six categories are described below:

(i) PeakInit: Papers whose citation count peaks within 5 years of publication followed by an exponential decay.

(ii) PeakMul: Papers having multiple peaks in different time periods of the citation history.

(iii) PeakLate: Papers having very few citations at the beginning and then a single peak after at least 5 years of the publication followed by an exponential decay in citation count.

(iv) MonDec: Papers whose citation count peaks in the immediate next year of the publication followed by a monotonic decrease in the number of citations.

(v) MonIncr: Papers having a monotonic increase in the number of citations from the very beginning of the year of publication till the date of observation.

(vi) Oth: Papers not belonging to any of the above mentioned categories belong to this category.

Figure 3 presents the temporal profile of average countX values for each of these 6 categories. Again, we can see that the average countX values are the highest for the $MonIncr$ and $PeakLate$ categories, which have been identified as having the categories corresponding to high number of citations in [7]. Similarly, average countX values are the lowest for the $MonDec$ and $Others$, which have been identified as the categories corresponding to the low number of citations (see [7] for details).

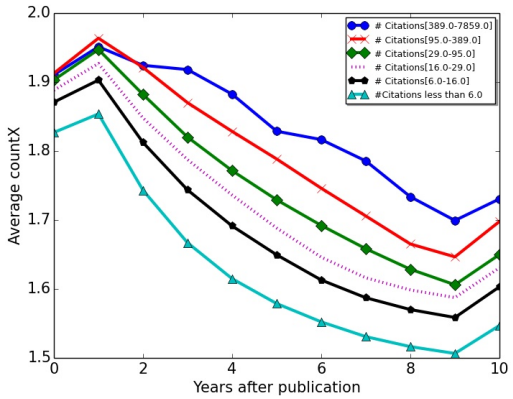

Figure 2: Average countX: temporal profiles for six citation buckets over the publication age

We now plot the temporal profile for the average citeWords values for the six citation buckets in Figure 4. Similar to average countX, while averaging for a citation bucket for a particular year, we consider only those papers which have a non-zero citation in that year. Average citeWords also shows a very similar trend as that seen with the average countX values, an initial increase and then a decreasing trend over the years. Interestingly, differences are observed between various citation ranges with the papers having the highest citations also earning a high number of average citeWords over the years.

We further use six citation categories to plot the temporal profiles in Figure 5. The trends are again very similar to those observed for the case of average countX values, with the MonIncr and PeakLate categories having a higher value of average citeWords than the other categories and MonDec category having the lowest values.