1. Discuss past and future government applications
2. Develop a more comprehensive related works section for Ch. 2
   (a) Setup context more effectively
   (b) More on general machine learning techniques
   (c) Discuss analogous research
3. Investigate big spikes in the time graph

   See Section 3.2

   Spikes represent Gmail folders being ported to the Hume center using an old script, which was responsible for changing the timestamps. A more sophisticated system was developed for transferring emails in Summer 2014, and there are no more spikes after this point. The biggest spike (from Jan 2014) came from moving large folders from a single graduate student. Based on the folder names, these are unrelated to the Hume center and were sent before this person was a graduate student, therefore these emails will be removed from the dataset for the purposes of this analysis.
4. Calculate measure of non-symmetry of Hume adjacency matrix

   See Section 3.3.2

   I measured the symmetry of the adjacency matrix using a metric I created because I could not find anything similar in the literature. I used a weighted averaging scheme so that larger email counts mattered more. On a scale where -1 represents perfect asymmetry and 1 represents perfect symmetry, I measured the adjacency matrix to have a value of 0.7036*. Let me know if you want more details on how I did this.*

   (a) Briefly explain the case of people emailing themselves - is it from CC-ed emails?

   See Figure 3.8. Yes, the majority of these emails did come from people copying themselves.
5. Look into mutual information values of zero and explain

   See Section 4.3

   There were four significant digits available for each value, so it did not appear to be a rounding error. There seemed to be an arbitrary cutoff point

around 0.22 bits. Lack of documentation makes it difficult to pinpoint the cause or correct for it, so I developed my own script to calculate the mutual information based on the description in Section 4.2. Table 4.1 has been updated to reflect the changes. Many of the features from the table were unchanged, but some are new.

6. Discuss a naive guess/baseline for classification

   See Section 5.2

7. Look into using a prior to represent the class distribution

   See Section 4.2

8. Add more commentary on LOOCV results

   See Section 5.2

   (a) Is there any source of data leakage? If not, explain disparity between classification and LOOCV results.

   (b) What types of people are misclassified?

   The types of people were misclassified were just what you would expect: phD students, postdoctoral researchers, new employees, etc. I think the most important point is realizing that within these 5 distinct classes, there exists both a range of functions and a range of personalities. I believe that the reason for the disparity between the classification and LOOCV results lies with the fact that the duties and behaviors can vary so much from person to person within a class. So much so that training on a person of the same class can lead to little or no information about the person under test. Having past information on a person to predict future behaviors allows the algorithm to recognize the past behavior as a sort of nearest neighbor (a high-performing classification technique from Figure 4.1) which vastly improves the classification accuracy.

9. For hierarchy analysis:

   See Section 5.3

   (a) Take direction into account for hierarchy connections

   I changed this analysis to only consider sent emails.

(b) Add more edges

I added edges (increasing in transparency) to represent 2nd and 3rd strongest connections. These changes are shown in Figure 5.7. *The more complex structure of the Hume center is now evident, but some of the divisions remain clear.*

10. Add collective classification to future work

See Section 6.2.2