Modeling Shared Drive Utilization Using Stochastic Techniques

Margret T. Martin, Sarah G. Nurre and Michael R. Grimaila, *Senior Member, IEEE*

***Abstract*—Information Technology (IT) units provide electronic file shared drives for utilization by personnel in their organization. This shared electronic storage space is used for a wide variety of reasons (e.g., archival, collaboration, backups, dissemination) and is generally focused on providing areas for collaboration, as well as to augment the primary storage disk space located within each user's computer system. The ways in which shared drives are utilized are highly dependent upon the organizational mission, who can access shared resources, the stability of the user population, end user roles, and the data retention policies enforced by the IT unit. The goal of this research is to understand what happens to information in shared disk storage within an academic institution as a function of time. Academic organizations are unique due to the transitory nature of the user population (e.g., students arrive and depart each year) and by the various roles that exist within the school. By examining the information lifecycle, we can gain insight into the differing perspectives between end users and IT units, the validity of assumptions about information rot and data aging, and develop an understanding how shared storage space is managed.**

In this paper, we evaluate the utilization of a file-share server used to manage official records within an academic organization and use Discrete Markov Chains to model and simulate the movement of stored data over time as a function of policy within an academic organization. The results show that different IT policies have a dramatic impact on the accumulation of information contained within shared storage space and that organizations should incorporate both the perspectives of the end users and the IT unit when developing organizational policies regarding the use of shared storage space.

***Index Terms*—Information Archival; Information Aging; Data Rot; Information Management; Records Management**

I. Introduction

Virtually all modern organizations have embedded Information and Communication Technologies (ICTs) into their core mission processes as a means to increase operational efficiency, improve decision quality, and reduce operational costs. Within an organization, the Information Technology (IT) unit is responsible for managing and maintaining the organizational ICT resources (e.g., computers, servers, and voice and data networks).

Academic organizations often leverage ICTs in support of the delivery of education to both in-residence and distance learning students, in support research activities, and to enable administration to be conducted in a cost effective, efficient manner. The daily activities of the administration, faculty, staff, and students of the modern academic institution require reliable and usable ICTs to properly attain their mission.

As such, the policies enforced by IT units when managing ICT resources within an academic organization are critical to mission success. One of the primary functions of the IT unit is the management of file-sharing servers, which enable the academic mission and facilitate collaboration within, across, and between institutional units and external collaborators.

End users within organizations have an insatiable demand for storage space. When new shared disk space is provided, inevitably it fills to capacity within a short period of time requiring management of quotas to be enforced so that a small number of users do not monopolize the shared resource. However, enforcing quotas on all users can interfere with the education and research mission of an academic organization when a user has a justifiable need for a large amount of temporary disk space and create a management burden to temporarily allow the user the required space. There are always tradeoffs between organizational policy, perceived user satisfaction, and management costs.

In this paper, we seek to answer the question "What happens to information stored on these shared drives as it ages from year-to-year?" within an academic organization. To answer this question, we first examine the utilization of shared disk storage space and then develop a discrete-time Markov Chain model used to simulate the evolution of stored information. The remainder of this paper is presented as follows. After a brief description and analysis of a shared storage drive within an academic organization under study and development of a conceptual model in Section II, we present the background of using Markov Chain models in Section III. In Section IV, we introduce a model that is used to simulate the impact that different information retention policies have on information storage over a five year period. Section V presents an analysis of the results, and is followed by concluding remarks in Section VI.

II. Conceptual Model Development

In order to accomplish the research objectives, we needed to develop a conceptual model of a shared storage drive. The first step in this model development was to examine the actual disk utilization of shared storage space used within an academic organization. We chose to examine only one of the many shared storage drives contained within the academic organization. The "official records" drive was chosen for analysis because of it clear purpose (i.e., it serves as a repository for the organizations official records) and the limited number of authorized users who have access to the drive. Table 1 below shows an overview of the official records drive and Table 2 below shows the summary statistics for this drive. Metadata describing the drive was collected by the IT unit using a Microsoft power shell script and subsequently inserted into Microsoft Access database to facilitate analysis.

Table 1. Official Records Drive Contents Overview

Total Number of Directories and Subdirectories	Total Number of Files	Total Space (Bytes)	Number of Unique File Extensions
96	49,885	44,530,209,647	365

Table 2. Summary Statistics for Official Records Drive

Mean	Standard Deviation	Minimum Size (Bytes)	Maximum Size (Bytes)
892657.30	5076508.94	2	334370852

There are several assumptions written below regarding the analysis of this data within the official records drive:

- Electronic files are counted based upon the number of unique extensions.
- Files are arranged based upon the calendar year the file was created. Even if a file was created in 2012, but placed on the shared drive/repository, in 2014, it counts as part of the 2012 data.
- File size is summed by year to determine the amount of storage space utilized.
- Data collection ended in May 2014, so the 2014 data point does not contain an entire year of data.
- It is important to realize not all data from each year is captured. This is because the IT unit clean-up policies subject each drive to certain deletion rates, based on when students graduate, preparations for inspections, and/or limitations pertaining to community drive space.
- The data is analyzed based upon the types of files users created, respective creation years, and is displayed using pie charts. This was accomplished by grouping files by category, which incorporated multiple extension types. These graphs provide a unique perspective of what was kept for archival purposes versus what was being created by the institution.

It is also important to note limitations of the analysis. Although the organizations maintain data for multiple years, the resources required to retrieve this data were not available. As a consequence, the data presented is a snap-shot of each drive's contents from May 2014. Ultimately, this was a single-point-in-time analysis. A request was made to the IT unit of the organization to provide the creator of each file; however the administrative support required to collect this data was not available. As a result, it was difficult to determine how many different users were contributing to the shared drive.

Our main intent in collecting data from the official records drive was to ask, "What is the organization working on and how are records preserving this as the spirit of the mission or as transparency in operations?" In order to understand where specific disconnects exists, we examined the data from a few different perspectives. We first sought to compare the growth of records in each of these file-sharing repositories to literature findings which state information grows at an exponential rate. We were interested to see if the collected data would support these findings. We hypothesized that while this may be true for individual user disks, we did not believe this would be true for a shared disk whose purpose was to house organizational official records. We suspected that population of the official records shared drive space would be driven by rate at which official records are created within the organization. Figure 1 below shows the cumulative number of bytes stored in the official records repository by year from 2005-2014.

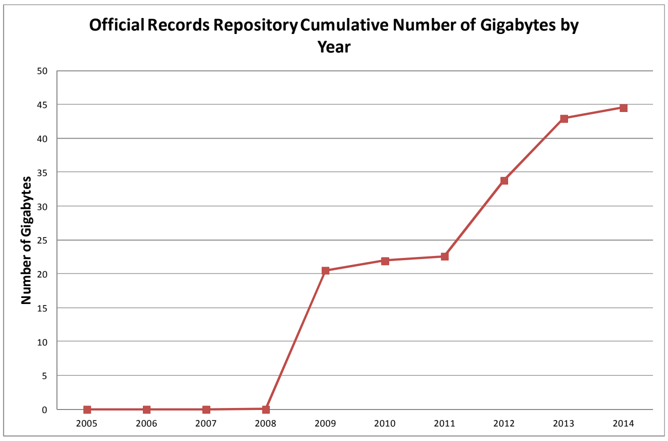

Figure 1. Official Records Repository (Drive) Cumulative Number of Gigabytes as a Function of Year

Notice that there are more dramatic increases in cumulative data in the period from 2008-2009 and 2011-2013. After consulting with Subject Matter Experts (SMEs), we determined that this can be explained by the fact that more data is added to the official records repository just prior to and during inspection years (2009 and 2012). Thus our belief that official records were added to the drive as they were created was incorrect. Instead, it appears that just prior to records management compliance inspections that the organization "rushes" to add files to the official records drive so that they will pass inspections. This analysis explains the organizational behavior with regard to the use of the official records shared drive. This is evident by the way the IT unit manages available storage space, the number and types of users, and the purpose of the drive. While it's possible to fit a trend line to the data, it would not add useful insight into understanding the growth of records in the official records repository because of its inherently piece-linear nature. Looking at the data from another lens, Figure 2 shows the files on the official archives drive by creation date. Notice the large number of files added in the years 2006, 2009, and 2012. This confirms the earlier observation that large numbers of files are added just before and during inspection years. Archiving and records management tend to be passive activities and the right resources and people must be available to conduct these efforts or else they will not occur until there is a tangible penalty (e.g., a failed inspection) levied against the organization.

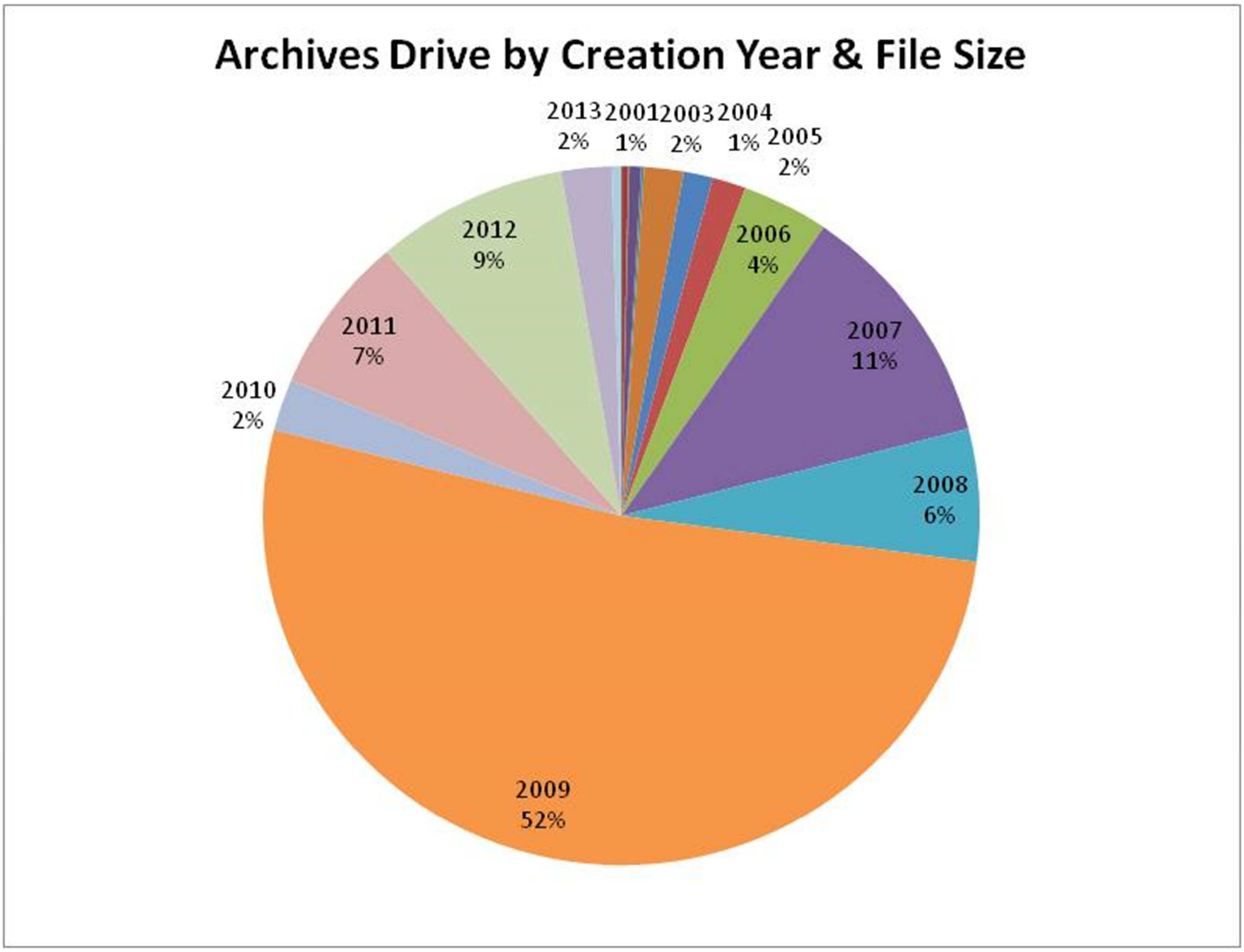

Figure 2. Official Records Repository (Drive) Files by Creation Year and File Size

We now examine the type of data stored on the official records drive. In this case, the number of files and their extensions were grouped by purpose and technological medium. For example, the "image" category encompasses file extensions such as .tif, .png, ,fig, .gif, .bmp, and .jpg. Note that only the top 25 extension types were considered for grouping purposes and the remainder was grouped into an "other" category. In the case of the official records repository, shown in Figure 3, the largest groupings consisted of 45% Adobe PDF files, 15% Microsoft Word documents, 7% computer languages, and 7% PowerPoint presentations. Only 15% were captured in the "Other" category. This was as expected, as the official records drive typically contains institutional records which are typically archived in the above common formats as opposed to general purpose drives that may contain a much broader spectrum of file types.

The analysis of data by extension types provided unique insight into what is important to the organization as a whole. A more in-depth analysis would have included grouping the information by user role, but unfortunately, the required data was not available.

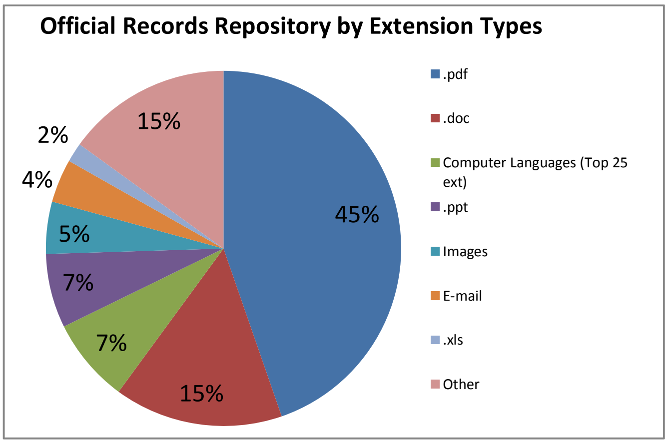

Figure 3. Official Records Repository (Drive) by Extension Type

IT governance principles explain how data is managed based upon who created it and what their role is within the organization hierarchy. For example, leadership files are seldom purged or deleted by the IT unit. In contrast, files created by students are regularly viewed for deletion.

The next step was to develop a generalized model to demonstrate an overview of information flow on the shared drive as shown in Figure 4. A time series analysis was conducted from the initial state of the drive in 2009, through the years 2010 through 2014. The initial input for this model system was labeled "Initial Info," and represents all of the electronic data including Word documents, Excel files, PowerPoint presentations that were created, posted, and last modified in, or prior to, the year 2009.

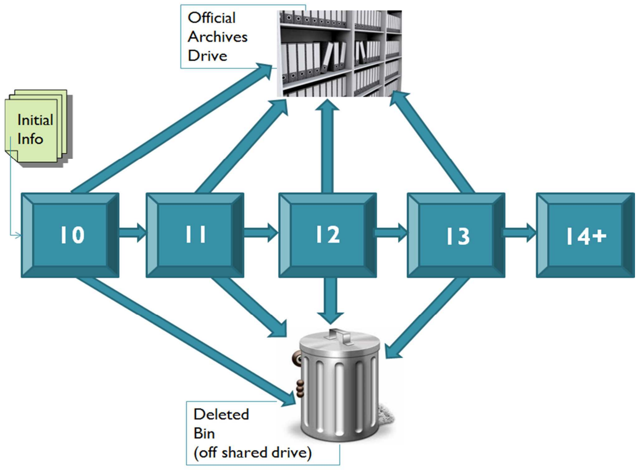

Figure 4. Generalized View of File-Sharing Server System

The model shown in Figure 4 is simplified and is used to evaluate what happens to "old information" as time increments from one calendar year to the next, given the possible paths this information can travel for the year range from 2010-2014. It is important to note that in this simplified model no additional information is added to the system in 2010 or thereafter. End users and IT Personnel in the academic user community have the following 3 choices:

1. Allow the initial data to stay on the shared drive (either for its inherent value, because a shared drive "clean-up" has not be conducted, or because users left it there). The option to keep "old information" is illustrated in Figure 4 by the horizontal lines running from one calendar year node to the next.
2. Archive the 2009 information onto the shared official records drive. Because the official records drive is not visible to a great majority of users, this is accomplished via the assistance of the academic institution's IT directorate and is depicted in the model by the arrows from each calendar year to the official records drive.
3. Purge the initial information. This option is demonstrated by the lines pointing to the trash receptacle from each calendar year node, which means it has moved off of the shared server space. Users, of course, are free to back-up their own information on personal storage devices as they choose and as allowed by the institution.

In a real world system, many issues stem from the fact that users add documents, data, and information to the system each year, but this difference is what allows this model to serve as a simple replication. Now that the basic, feasible paths are explained, the next step is to apply stochastic modeling concepts.

III. Markov Chain Models

Previous research has been conducted using Markov Chains in order to evaluate ICT phenomena. For example, Yossef et al. [1] used a one-dimensional Markov Process, or random walk, to approximate certain aggregate queries, such as search engine usage and the proportion of pages belonging to .com or other domains. Attempts to estimate the size of a domain, or estimating the fraction of web pages covered by a search engine are both efficient and require very limited resources. Thain et al. explained that end users and systems administrators have "two distinct roles to play" and the importance of IT professionals being able to apply set constraints while users must be given elements of freedom to work as their mission requires without extreme limitations or constraining policy requirements [2]. This can involve the implementation of distributed storage systems with two distinct intents, or services: storing data vs. organizing directories. Ultimately, this research highlights the idea that administrators shouldn't care about the purpose for why a user is employing a file server, with the exception of security reasons and resource policies. Flexible policies should be set in place to lead to new modes of interactions for users.

The distinction between the interpreted value (or usefulness of the data) vs. an IT administrator's due diligence in managing limited storage, server space can lead to some interesting assumptions from both ends of the spectrum. Whether it's accidental or intentional deletion of data, it's important to realize that any risks and faults, albeit latent or visible, are memoryless according to Baker et al, and similar to a Markov Chain [3]. Additionally, two "dangerous assumptions" that the article mentions are an unlimited budget assumption and human error which are disconnects when conduction long-term digital preservation. Work is this realm is in high-demand. This is evident by works such as Fessant, et al who specifically recommend further analysis of peer-to-peer networks [4] and Z. Ge et al [5], as well as research pertaining to cost-effective file migrations of servers [6].

IV. A Discrete Markov Chain Model

To stochastically model this system, we developed a discrete-time Markov Chain. A Markov Chain is a mathematical system that undergoes transitions from one state to another and is deemed 'memoryless' such that what happens in the future depends solely upon the current state and the probabilistic determination of the projected path.

Note that in Figure 5, the model was updated so that the official records drive (the O:/ drive) now appears to 'recycle' back onto the calendar year nodes. Modifying the archival option for the data to a transient state at each calendar year node is important because the official archives drive does count towards the institution's storage limit authorized, in terms of shared server space and it shows that information which was once archived can be moved back onto the drive so it will continue to the next calendar year or it is moved to the deletion bin. This is an important caveat pertaining to the institution's records management regulations. The only two absorbing states in this Markov Chain are the 2014+ node and the deleted items bin. Once a transition into an absorbing state occurs, the information, or data, will remain there forever.

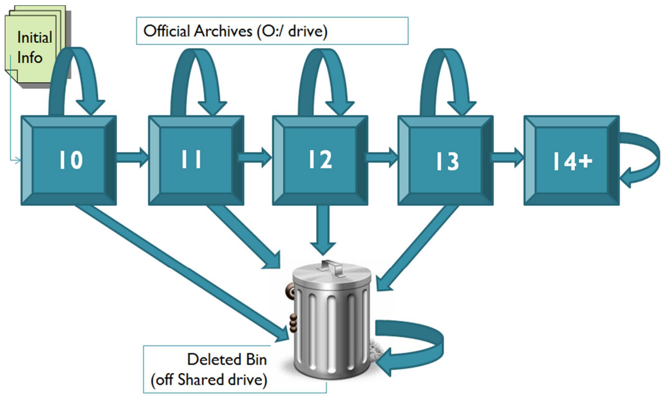

Figure 5. Markov Chain Model of Information Flow

A. *Assumptions*

Because longitudinal data was unavailable for the official records drive, a table of important perspectives and assumptions was devised which led to an analysis stemming from two different lenses by which to view the shared drives as shown in Table 3 below.

Table 3. Information Retention Policy Perspectives/Assumptions

Perspectives	IT Perspective	Only so much storage space is available in the institution; IT's job is to archive official records & manage shared drive space (but not to determine the value of the information).
	Users Perspective	There is value added by keeping information (i.e. mission requirements, "just in case we need it"); Generally purge information only when prompted to do so.
Assumptions	General	A large percentage of information "rots" with time (especially when it is not updated or utilized).
	Even-numbered Years	Records Management inspections usually occur during even-numbered years; IT is more apt to archive e-files on the official archives drive then.
	Odd-numbered Years	Server space fills up on these years, so IT initiates 'clean-sweeps' to encourage users to purge unnecessary information.

B. *Analysis*

After the academic institution's IT and user perspectives were realized, two Markov Chain transition matrices [7] using randomized probabilities were qualitatively generated using SME inputs. These values had to be assumed due to the lack of availability of the time series data required to properly estimate these parameters. Table 4 shows the probability transition matrix representing the IT personnel perspective and Table 5 shows the probability transition matrix representing the academic user perspective. The models were verified by multiple sources within the IT realm, including a PhD with a background in Information Sciences.

As an example, consider Table 4 which shows the P-matrix from the IT personnel perspective. In this case, the percentage of 2009 information that is likely to stay on the server from 2010 to 2011 is 86%, while 93% of this data would be retained if users exercised their organizational behavior. The transition matrices provide estimated probabilities associated with the transition from one state, or calendar year (the headings on the left of the matrix) to another (the headings along the top of the matrix).

In both matrices, the far right hand column shows the likelihood of information being deleted off the file-shared drive. The IT unit's philosophy is based on the idea a specified amount of storage space exists, so this directorate must actively delete electronic records after a certain time period assuming the concept of information rot. Users do not abide by this same rationale, and instead, only exert clean-up efforts when prompted to do so, by IT personnel working with the institution's leadership.

Table 4. P-Matrix for IT Personnel Perspective

\To From\	10	11	12	13	14	Delete
10	0.04	0.86	0	0	0	0.1
11	0	0.02	0.7	0	0	0.28
12	0	0	0.03	0.55	0	0.42
13	0	0	0	0.02	0.43	0.55
14	0	0	0	0	1	0
Delete	0	0	0	0	0	1

Table 5. P-Matrix for Academic User Perspective

\To From\	10	11	12	13	14	Delete
10	0.02	0.93	0	0	0	0.05
11	0	0.02	0.9	0	0	0.08
12	0	0	0.01	0.94	0	0.05
13	0	0	0	0.01	0.91	0.08
14	0	0	0	0	1	0
Delete	0	0	0	0	0	1

The numbers along the primary diagonal of each matrix show the probability of archival, which are slightly elevated during records management inspections years, especially because the IT unit owns and is responsible for the records management program. The remaining positive values, all less than 1.0, show the amount of 2009 information that moves from one calendar year to the next on the shared drive. Notice by the lack of participation in cleaning and maintaining the drives and information, that users have a much higher probability of letting 2009 information stay on the drive as opposed to removing it. Using occupancy probability matrix equations [8], the aforementioned transition matrices (i.e. P-matrices) were evaluated to calculate the π values and n-step transition matrices associated with the number of time-steps to move all the initial information from 2009 into the absorbing states. Figure 6 shows the simulated information retention rates as a function of time for years 2010-2013 (corresponding to n=2, 3, 4, and 5) for IT operations personnel when using the P-matrix shown in Table 4. Figure 7 shows the simulated information retention rates as a function of time for years 2010-2013 for academic users when using the P-matrix shown in Table 5.

n=2 2010
[0.0016 0.0516 0.602 0.0 0.0 0.3448]
n=3 2011
[0.0 0.0024 0.0542 0.331 0.0 0.6122]
n=4 2012
[0.0 0.0001 0.0033 0.0397 0.142 0.8178]
n=5 2013
[0.0 0.0 0.0 0.0 0.158 0.8392]
Keep Delete

Figure 6. Information Retention Rates for IT Operations Personnel

n=2 2010
[0.0004 0.0372 0.837 0.0 0.0 0.1254]
n=3 2011
[0.0 0.0011 0.0419 0.7868 0.0 0.1702]
n=4 2012
[0.0 0.0 0.0014 0.0472 0.7160 0.2354]
n=5 2013
[0.0 0.0 0.0 0.0018 0.7589 0.2392]
Keep Delete

Figure 7. Information Retention Rates for Academic Users

Evaluating time steps, from 2010 through 2013, simulates the movement of information on the shared drive from year to year and reveals the information retention after 4 years (at n=5) for the official archives. Assuming the IT and user 'policies' ran their courses separate from one other, this analysis demonstrates that after 6 years, the IT Directorate would delete approximately 83.92% of the 2009 data off the shared drive and 15.8% would still remain on the drive. Relying on a "users decide" policy, 75.89% of the initial 2009 information would still remain on the shared drive while only 23.92% would be deleted.

V. CONCLUSIONS

The intent of this research was to better understand what happens to information on a shared drive as it ages and study the amount of information that accumulates by evaluating differing policies, or perspectives. In order to scope this project appropriately and remain sensitive to the workload required by the institution's IT unit, interviews were conducted with the IT Director and various SMEs. A discrete-time Markov Chain was created and an analysis was conducting to determine what the anticipated information flow looks like from year-to-year.

The analysis behind this discrete-time Markov Chain and associated probability matrices demonstrates a genuine disconnect in the way users and IT personnel view and treat shared server space based on the keep-to-delete ratios. When using the IT operations personnel policies, the keep-to-delete ratio was approximately 16:84. In contrast, when using the academic user's policies we obtain a keep-to-delete ratio of 76:24 over the same time period. This is not surprising and notionally matches the behavior observed in the organization. Users will store files perpetually, if allowed, so that they will have another backup of their critical records.

While both perspectives have valid viewpoints, the analysis behind this discrete-time Markov Chain and associated probability matrices demonstrates a genuine disconnect in the way users and IT personnel view and treat shared server space based on the keep-to-delete ratios. Because so much untouched, unreferenced, and unmodified information has a way of accumulating on these servers, it is no wonder there is seldom enough storage space available at this academic institution and that the shared drive icons turn red so frequently, signaling they are close to maximum capacity as seen in the drive utilization shown in Fig. 8.

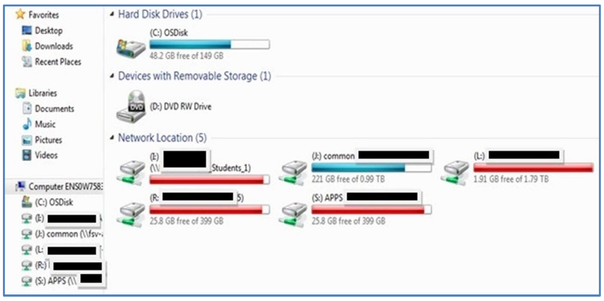

Figure 8. Academic Shared Server Space Utilization by Drive

It is recommended that additional operations research concepts be applied to this research, such as Markov Decision Processes and Bonus-Malus systems which can allow differing costs and incentives to be associated with the decisions to better understand organizational behavior, trends, and related policies. As previously discussed, even if the IT policy determines that 16% of old data from 6 years prior should remain on the storage drive, and assuming that same amount of information is retained year after year, nearly 20% of the server will be comprised of data 5+ years old. Naturally, this is just hypothetical, but would not be feasible if the amount of storage space ceases to increase.

Many of the assumptions in this paper derive from the mathematical probabilities related to human behavior and the older information and data becomes, the higher probability it has to become archived or deleted based on obsolescence. In addition, it's important for follow-up research to be conducted to truly evaluate which files are used, accessed, and modified vs. that which is retained and never retrieved, which may include many of the archives. An important follow-on question is: Can 'carrying costs' be associated with the movement of information on a file-sharing drive in order to create a more effective policy?

VI. Disclaimer

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the U.S. Government.

References

[1] Z.B. Yoseff, A. Berg, S. Chien, J. Fakcharoenphol, D. Weitz. (2000.) Approximating Aggregate Queries about Web Pages via Random Walks. Proceedings of the 26th VLDB Conference, Cairo, Egypt. pp 535-544. [Online]. Available: http://www.vldb.org/conf/2000/P535.pdf

[2] D. Thain, S. Klous, J. Wozniak, P. Brenner, A. Striegel, J. Izaguierre. (2005, Nov.) Separating Abstractions from Resources in a Tactical Storage System. Seattle, Washington. [Online]. Available: https://www3.nd.edu/~ccl/research/papers/tss-sc05.pdf

[3] M. Baker, M. Shah, D. Rosenthal, M. Roussopoulos, P. Maniatis, T. Giuli, and P. Bungale. (2005, Aug.) A Fresh Look at the Reliability of Long-Term Digital Storage. ACM SIGOPS Operating Systems Review. Vol. 40. No. 4. pp 1-14. [Online]. Available: http://arxiv.org/pdf/cs/0508130.pdf

[4] F.L. Fessant, S. Handurukande, A.-M. Kermarrec, L. Massoulie. (2004.) Clustering in Peer-to-Peer File Sharing Workloads. IPTPS pp. 217-226

[5] Z. Ge, D.R. Figueiredo, S. Jaiswal, J. Kurose, D. Towsley. (2003). Modeling Peer-Peer File Sharing Systems. *Proceedings of IEEE INFOCOM.* pp. 2188-2198.

[6] B. Gavish and O.R. Liu Sheng. (1990, Feb.) Dynamic File Migration in Distributed Computer Systems. Management of Computing, Vol. 33, No. 2, pp. 177-189.

[7] Sheldon M. Ross. *Introduction to Probability Models, 9th edition.* Academic Press, 2006.

[8] C. Richard Cassady, J.A. Nachlas. Probability Models in Operations Research. (2008, Aug.) CRC Press.