Understanding Data Sharing in Private Personal Clouds

Raúl Gracia-Tinedo and Pedro García-López
Universitat Rovira i Virgili, Tarragona (Spain)
{raul.gracia|pedro.garcia}@urv.cat

Alberto Gómez and Anastasio Illana
NEC, Madrid (Spain)
{alberto.gomez|anastasio.illana}@emea.nec.com

***Abstract*—Data sharing in Personal Clouds blurs the lines between on-line storage and content distribution with a strong *social component*. Such social information may be exploited by researchers to devise optimized data management techniques for Personal Clouds. Unfortunately, due their proprietary nature, data sharing is one of the least studied facets of these systems.**

In this work, we present the first study of data sharing in a private Personal Cloud. Concretely, we contribute a dataset collected at the *metadata back-end* of NEC: an enterprise oriented Personal Cloud. First, our analysis provides a deep inspection of the *storage layer* of NEC, comparing it with a well-known public vendor (UbuntuOne). Second, we study the *social structure* of NEC user communities, as well as the storage characteristics of user *sharing links* via multiplex network techniques.

Finally, we discuss a battery of *data management optimizations* for NEC derived from our findings, which may be of independent interest for other similar systems. Our proposals include content distribution, caching and data placement. We believe that both our study and dataset will foster further research in this field.

***Index Terms*—Cloud Storage; Personal Clouds; Social Network Analysis; Measurement**

I. Introduction

Today, Personal Cloud services are becoming a widespread technology among users and organizations to *store*, *synchronize* and *share* files [1], [2], [3], [4]. Due to its flexibility, the Personal Cloud model has been successfully materialized in different forms; while most users are attracted by well-kown public providers, such as Dropbox and Box, enterprises also resort to private or hybrid Personal Cloud solutions (e.g., ownCloud[1], StackSync[2]) to retain the control of their data.

From a Personal Cloud viewpoint, to satisfy ever-increasing efficiency and scalability service requirements calls for devising advanced data management techniques [5]. To wit, desktop clients contact the server infrastructure to upload and/or download files, or even to modify already stored ones. Behind the scenes, the data management activity of a Personal Cloud can be complex, involving data compression, deduplication and content distribution techniques, among others [6], [5], [7].

Despite the associated technical complexity, events like the recent shut down of UbuntuOne evidence that Personal Clouds need to understand the activity of users and optimize data management accordingly [8], [2]. Otherwise, vendors might suffer excessive storage costs related to their daily workloads.

In this sense, one of the services that accounts for an important fraction of data management activity is *data sharing* [1]. That is, a major promise of a Personal Cloud service is enabling users with easy and integrated *data sharing* capabilities. Within the service domains, users can share data in various ways (e.g., public download link, native sharing) across a variety of devices and operating systems (OSes).

Personal Clouds blur the lines between a storage and a content distribution service with a strong *social component*. That is, sharing interactions within the system are driven by the work or social links of users. In our view, to analyze the social component of Personal Clouds will provide important insights of how users share data in these systems. Even more, such information may be key to derive novel data management optimizations or to effectively tailor existing ones.

Unfortunately, while previous research works targeted other technical aspects of public Personal Clouds (e.g., protocols, storage management) [6], [5], [7], data sharing is perhaps one of the least studied facets of these systems. The main reason is that analyzing data sharing interactions in a Personal Cloud with high resolution requires accessing the metadata of the service back-end. However, given the proprietary nature of most services, such a kind of analysis is not always possible.

To fill this gap, in this work we analyze the sharing activity of NEC[3]: a private, enterprise-oriented Personal Cloud. The NEC Personal Cloud supports hundreds of enterprises and thousands of users/employees making regular use of the service. Concretely, we provide an analysis of the *storage layer* and *sharing activity* of NEC thanks to the metadata collected by the provider itself. Such a unique dataset has allowed us to study the sharing activity of a private vendor in a high level of detail. Further, we can also compare characteristics of NEC with other public vendors. In summary, our contributions are:

- *Storage and user activity analysis*: First, we focus on understanding the characteristics of files stored and shared in NEC compared with a public vendor (UbuntuOne), as well as the activity of users.
- *Social network implications*: We analyze the NEC data sharing network, including its topological characteristics, the clustering among users and the relationship between user connectivity and activity in the system.
- *Multiplex network analysis*: We also resort to multiplex

[1] https://owncloud.org/

[2] http://stacksync.org/

[3] http://www.nec.com/en/global/solutions/cloud/portfolio/storage.html

network techniques to correlate the different networks and user roles emerging from defining different types of sharing links across users (e.g., content types, intensity).

- *Potential optimizations for Personal Clouds*: From our analysis, we discuss data management optimizations for NEC-like systems, such as: i) Socially-informed caching for highly shared files, ii) peer-to-peer file synchronization for clustered user groups, or iii) topology-based data placement at the server infrastructure, among others.
- *Publicly available dataset*: Finally, the collected dataset is publicly available online[4]. We believe that this dataset will help to foster further research in this field.

The rest of this paper is structured as follows. We discuss the related work in Section II. Our analysis methodology and dataset is described in Section III. The analysis of data sharing activity in NEC encompasses sections IV, V and VI. In Section VII we discuss data management optimizations derived from our findings. We conclude in Section VIII.

II. Related Work

Recently, Personal Cloud services have attracted the attention of the research community due to their broad adoption and commercial popularity. The seminal work of Drago et al. [1] presented an external measurement of Dropbox in both a university campus and residential networks. They analyzed the traffic generated by users, as well as the workflow and architecture of the service. Similarly, other works tried to infer the closed architecture and operation of Personal Clouds, such in the case of Wuala [3] and UbuntuOne [2].

As a motivation for this work, we found particularly interesting research efforts devoted to improve data management in Personal Clouds. Li et al. investigated the problem of frequent file modifications in the synchronization client of Dropbox [6], proposing a solution to this data management problem. Authors in [5], [7] extended this effort and provided a comprehensive analysis of the data management techniques integrated in desktop clients (e.g., compression, deduplication, etc.). These works show the importance understanding the behavior of users to optimize data management in this setting.

However, none of the previous works proposed data management techniques that benefit from social information related to data sharing. The main problem is the absence of data sharing characterizations in the Personal Cloud arena.

To our knowledge, the only attempt on measuring data sharing in Personal Clouds is [9]. In this work, Gonçalves et al. characterize data and metadata flows of Dropbox desktop clients related to shared folders in a campus. However, the external and "black box" approach adopted in [9] limits the insights of this study. That is, authors did not provide information about the social topology of Dropbox users in the campus, a characterization of sharing links or the files that users share, to name a few. This work overcomes these limitations thanks to the back-end information of our dataset.

Our analysis of data sharing in Personal Clouds is partially inspired by existing works of social network analysis [10], [11]. Social network analysis enables us to understand topological characteristics of user communities. In particular, we found specially interesting the work of Iamnitchi et al. [12] that examines in depth the data sharing network of different client-server and peer-to-peer systems. However, we go a step further by applying multi-layer or multiplex network analysis [13], [14], [15] to illustrate how users are interconnected depending on different link types (e.g., intensity, content type).

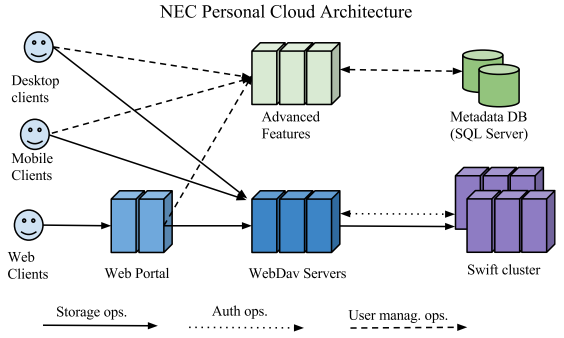


Fig. 1. Architecture of NEC Personal Cloud back-end (Madrid datacenter).

Moreover, our study is the first to compare the storage layer of a private vendor (NEC) with a public one (UbuntuOne). The new insights provided by our analysis enable us to discuss the application of data management techniques in Personal Clouds (client/server sides) to motivate further research in the field.

III. Background and Methodology

In this section, we provide the necessary background to facilitate the understanding of our study. This includes a description of the architecture of NEC, as well as a definition of the metrics and methods used in our analysis. Moreover, we provide a description of the collected dataset.

A. Overview of the NEC Personal Cloud

As other Personal Clouds [1], [2], [16], NEC[5] exhibits a 3-tier architecture consisting of *metadata service*, *storage back-end* and *desktop or mobile/web clients* (see Fig. 1).

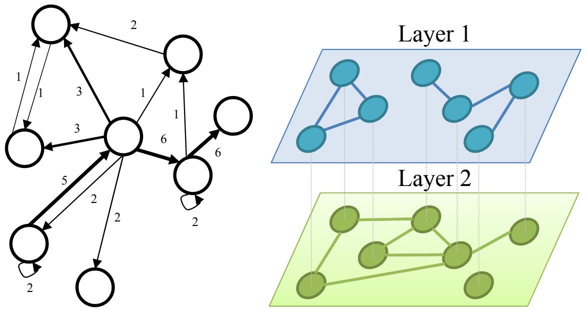
In the back-end of NEC, we find the metadata service and the data store. On the one hand, the metadata service of NEC is composed of 3 elements: WebDav servers, Advanced Features servers and Web portal. Similarly to ownCloud[6], NEC relies in WebDav [17] as a protocol for providing remote file management. These servers receive client requests related to the management of files, such as creating or deleting files. The Advanced Features service in NEC carries out relevant tasks to its internal operation, such as maintaining the logical relationships among users and organizations or tracking the storage quota of users. However, the most important function of this service is to manage sharing relationships of users; sharing links are persistently stored in a database (SQL Server). Finally, the Web portal provides users with means for managing their files via a regular Web browser.

On the other hand, the actual file contents of users are stored in an OpenStack Swift cluster, which offers simplified storage management and high scalability. In fact, other services such as Dropbox and UbuntuOne are known to use object storage to store user files [1], [2], [16]. As shown in Fig. 1, the WebDav

[4] http://cloudspaces.eu/results/datasets

[5] NEC Personal Cloud deployment at Madrid (Spain) SaaS facilities.

[6] http://owncloud.org



(a) Users (nodes) and sharing interactions (edges or links) described as a directed graph.

(b) Example of multiplex network.

Fig. 2. Graph modeling in our social network analysis.

servers access Swift to serve user data operations. Moreover, NEC delegates authentication service to OpenStack KeyStone.

Outside the datacenter, NEC desktop clients are perhaps one of the most popular ways of accessing the service. They are client-side software that interact with the server infrastructure via a synchronization protocol that defines data and metadata operations. NEC also provides alternative access to the storage service, such as mobile and Web clients.

B. *Network Analysis Concepts and Metrics*

First, let us describe a social network as a directed graph $G(V,E)$, where V is the set of users represented as nodes and E the set of directed links that represent their sharing interactions. We define a directed link as $v_i \overset{w}{\to} v_j$ to express that user v_i *shared* w files with v_j. In our analysis, this model facilitates to distinguish between producers and consumers of shared content, as well as the intensity of links (see Fig. 2a).

The *connectivity or degree* of a node v, k_v, is the number of attached edges. As we model the social network as a directed graph, we distinguish between *out* and *in* degrees. Moreover, as NEC enables users to share data across its own devices and accounts, we also consider *self* links.

To measure clustering, we resort to both the *local clustering coefficient* and *the network clustering*. The former metric, namely C_v, is the ratio of the number of existing links over the number of possible links among its neighbors. For directed graphs, it is defined as: $C_v = e_v/(k_v(k_v-1))$, where k_v is the number of neighbors of v and e_v is the number of connected pairs among all neighbors of v. Alternatively, we use the network clustering coefficient (C) that represents the mean of local clusterings: $C = \frac{1}{|V|}\sum_{v \in V} C_v$.

We also resort to basic multiplex network analysis [13], [14], [15]. In a nutshell, multiplex or multi-layer analysis describes *multiple link types among the same set of nodes* within a network. As shown in Fig. 2b, graphs emerging from defining different link types (e.g., docs, pictures) among nodes are embedded into separate layers. In this paper, multiplex analysis helps us to understand the similarity or overlap across network layers (links, degree). Thus, we first use the *link overlap* between two layers $G_i(E_i,V_i)$ and $G_j(E_j,V_j)$ expressed by the Jaccard coefficient ($\mathcal{J}$) as $\mathcal{J} = \frac{|E_i \bigcap E_j|}{|E_i \bigcup E_j|}$. We also use the *degree overlap*, given by the Pearson correlation coefficient between the node degrees in pairs of layers.

Users or nodes	$8,393$
Connected components	$1,034$
Sharing links or edges	$43,244$
Sharing trace period	803 days
Files in OpenStack Swift	$76,645$
Storage consumption	1.96TB

TABLE I
DATASET SUMMARY.

C. *Dataset*

The NEC dataset integrates two sources of information: *storage layer* and *sharing interactions*. The information to build our dataset was collected directly by the provider from the back-end (SQL Server). A summary of both types of information sources can be found in Table I.

Regarding storage, the trace is a snapshot of the data store contents (OpenStack Swift). To wit, the trace contains log lines that identify and describe files (size, extension), as well as the file owner and the container/folder where it is stored. This enables us to analyze the storage layer of this service in detail.

The sharing trace contains log lines describing sharing interactions across users and information about shared files. This trace contains all sharing interactions in the NEC Personal Cloud (Madrid datacenter) from March 7th 2013 to September 9th 2015. To our knowledge, this is the most extensive trace of data sharing in Personal Clouds to date.

Dataset limitations: The collected dataset has some limitations. For instance, we did not trace information about the locations of users (e.g., IP addresses) due to privacy reasons. Moreover, it should be noted that while the sharing trace captures the whole sharing activity of users for several months, the storage layer information corresponds to a snapshot at a specific point in time. Consequently, not all the files captured in the sharing trace exist in the storage trace, and vice-versa. The sharing trace also lacks from temporal information.

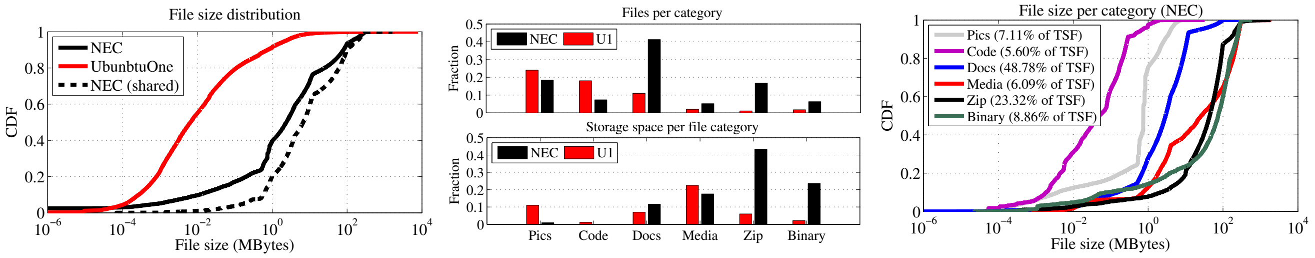
IV. BEHIND SHARING: STORAGE LAYER ANALYSIS

Next, we analyze the files stored and shared in NEC. From a provider's viewpoint, we believe that such analysis is essential to understand the implications of sharing in terms of data management. For the sake of clarity, we compare NEC with UbuntuOne (U1), the Personal Cloud of Canonical [2]. This will help us to highlight usage differences between both services, giving generality to our results (see Fig. 3).

A. *Snapshot of Storage Layer*

In Fig. 3a, we compare the sizes of files in NEC and U1. As we can observe, NEC files are in general significantly larger than files found in U1. To inform this argument, whereas the 90% of U1 files are smaller than 1MB, in NEC this value is 80.15MB. In terms of data sharing, this observation has a relevant implication: sharing and synchronizing large files across multiple clients *consumes more server resources* than for small files (e.g., upload bandwidth).

Fig. 3a also shows the size of files stored in NEC that have been object of sharing, which represent 26% out of the total file ids in the trace; clearly, NEC users exhibit a tendency to *share larger files*. Thus, given the same rate of sharing activity, NEC would have to deliver a higher amount of upload bandwidth compared to U1. Despite this may have economic repercussions for the provider [8], the fact that



(a) File sizes of NEC vs UbuntuOne.

(b) Relative storage space and popularity per file category (NEC vs UbuntuOne).

(c) Sizes of stored files per category in NEC (TSF=Total Shared Files).

Fig. 3. Analysis of shared and stored files in the NEC Personal Cloud.

users share larger files opens the door to effectively exploiting *content distribution techniques in desktop clients* to minimize bandwidth consumption [18].

To explain why files in NEC are larger than in U1, we aim at inspecting the dominant types of files that exist in both services. Thus, we classified files related to the 80 most popular file extensions according to six categories: pictures (*Pics*), code files (*Code*), documents (*Docs*), audio/video files (*Media*), compressed files (*Zip*) and binary/executable files (*Binary*). Fig. 3b shows the relative popularity and storage spaced consumed by these categories in NEC and U1.

Appreciably, in Fig. 3b the dominant file types in NEC and U1 differ considerably. In NEC, the most frequent files are documents, whereas in U1 the most common file types are code files. Moreover, conversely to U1, one of the most common file categories in NEC is *Zip*. This evidences that users manage *different data types depending on the usage* they exhibit in Personal Cloud service at hand.

Considering the storage space of file categories, in NEC compressed, binary and media files dominate 85% of the total storage consumption. If we consider that most of these files are *hardly compressible* [19], this yields that applying data compression to minimize outgoing bandwidth in data sharing *may be ineffective*. Similarly, in U1 *Media* and *Pics* categories account for the largest fraction of storage space, which also represent files with hardly compressible contents.

The reason why storage space is more dominated by non-compressible files is due to their size. In Fig. 3c we observe that *Zip*, *Binary* and *Media* files are much larger than code files or documents. In terms of sharing, 48.78% of the files shared in NEC correspond to documents. This is not surprising given that the NEC Personal Cloud was mainly used in corporate environments. Fig 3c also shows that *Zip* files not only account for a large fraction of storage space, but also for a significant amount of sharing activity (23.32% of total shared files).

Insight 1: *The nature of contents (file types, sizes) differs across Personal Clouds, depending on the habits of users. This impacts on the resources required to provide data sharing.*

Insight 2: *Larger and incompressible files account for a majority of the storage space in NEC compared to UbuntuOne.*

B. User Storage and Sharing Activity

Next, we analyze the consumption of resources across users. We identify how users behave, both in terms of storage and sharing activity. To this end, in Fig. 4, we differentiate the sharing activity of a user v_i to another user v_j based on: i) if user v_i shares files with user v_j ($v_i \rightarrow v_j$, *"sharing out" link where v_i is a "producer"*), ii) in the opposite case ($v_j \rightarrow v_i$, *"sharing in" link where v_i is a "consumer"*), and when v_i shares files with himself ($v_i \rightarrow v_i$, *sharing self*) —for example via a sharing hyperlink. We also show the files stored per user.

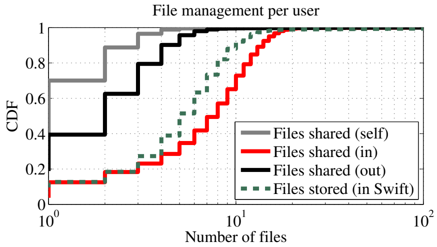


Fig. 4. Number of stored and shared files per user.

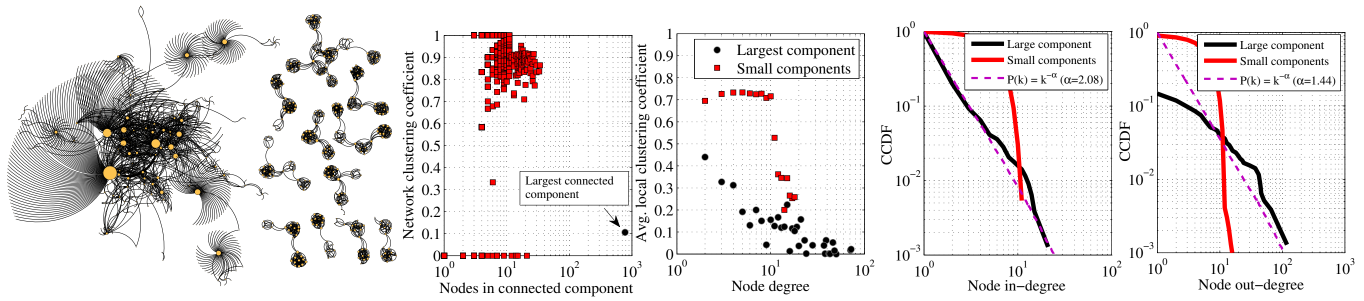
In Fig. 4, we observe that most NEC users are *consumers of shared data*. That is, only 3.7% of NEC users have not requested any shared file —whereas 50% of users requested 8 or more files. Conversely, the amount users that produce shared content is more limited; for instance, only 1% of users shared 8 or more files with others. The self-sharing activity is even less important than the production of shared contents.

We also infer from Fig. 4 that storage is an important aspect of the activity of users —e.g., 90% of users stored some content in NEC. Moreover, similarly to what happens with sharing activity, the *storage consumption of users is skewed*; i.e., 10% of most active users account for 38.55% of storage consumption. Activity skewness has been also observed in UbuntuOne [2], which seems to be an orthogonal phenomenon of these systems. This opens the door to apply specific storage optimizations —or even limitations— on very active users.

Perhaps surprisingly, we have not found correlations among the storage and sharing activity of users. That is, the Pearson correlation coefficient comparing the number of shared (self/in/out) and the number of stored files per users falls below ±0.3, indicating no correlation. We also looked for correlations among the different types of sharing obtaining similar results. This means that the behavior of NEC users *is specialized or stereotyped*; users may be intensive at either storing files or at some form of sharing (consumers, producers), but normally users do not exhibit the same intensity in more than one type of activity.

Insight 3: *Most NEC users tend to store and consume shared data; producers of shared content are a minority.*

Insight 4: *The activity of NEC users is skewed and stereotyped.*



(a) Largest and (some) smaller connected components in the NEC data sharing network.

(b) Network clustering coefficient vs component size (left) and avg. local clustering vs node degree (right).

(c) Complementary CDF of node degrees in the largest/smaller components.

Fig. 5. Topology characteristics of the NEC data sharing network.

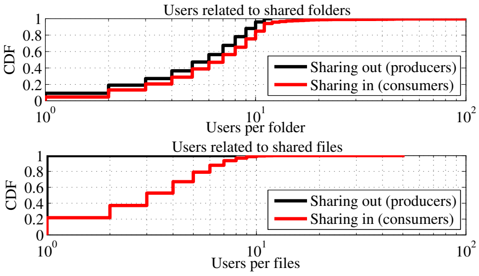


Fig. 6. Number of consumers and producers of shared files and folders.

C. Inspecting Shared Files and Folders

In what follows, we aim at inspecting the number of collaborators (producers, consumers) that interact with shared files and folders (see Fig. 6).

On the one hand, Fig. 6 shows that *most shared folders exhibit a similar number of producers and consumers*. However, the tail of the distributions is where the actual differences arise; whereas the percentile 99.9% of shared content producers per folder is 11, in the case of shared content consumers this number is 138. Thus, there is a small number of shared folders that represents a *hot spot of sharing and storage activity*. Considering the locking mechanisms of WebDav for shared resources [17], NEC shared folders with many concurrent users may incur in performance degradation.

Files are almost always shared by a single producer and to multiple consumers. Similarly to the case of folders, the long tail of the file consumers distribution points out that a small fraction of files is shared with a high number of users. Such a *skewed file popularity* is common in other storage domains and it can exploited via caching techniques; accordingly, Personal Cloud desktop clients may integrate socially-informed caching techniques —specially within LAN corporate environments— to reduce transfers with the datacenter [20].

Moreover, we have not found correlations between the number of shared content producers per folder and the amount of files or the storage space of the folder. However, if we relate the number of consumers of a shared folder and its storage space we obtain a weak correlation —Pearson correlation coefficient is 0.67. This yields that, in some cases, the storage space of a shared folder is related to the number of consumer users.

Insight 5: *Shared files and folders exhibit a skewed distribution of consumer users.*

V. Data Sharing Social Network

Next, we analyze the structural characteristics of the NEC Personal Cloud data sharing network (see Fig. 5). Such analysis will help us to understand how users share data across them, which is in turn a valuable source of information to develop advanced data management techniques for these systems.

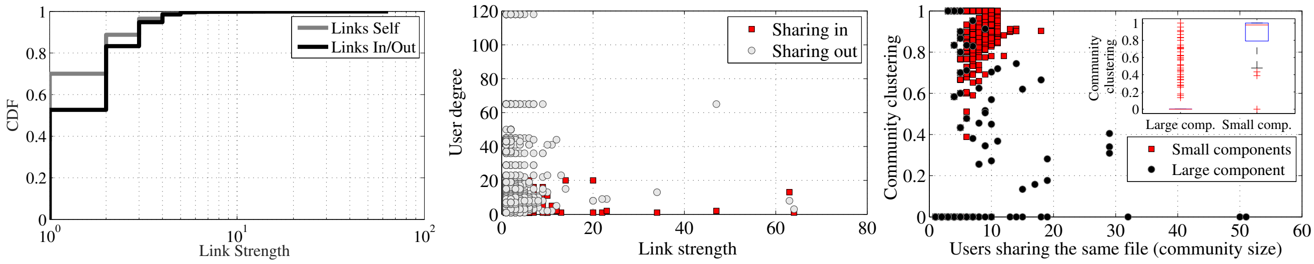
A. Network Analysis & Characterization

NEC users are not connected to a single network, but rather distributed across many connected components that coexist within the same Personal Cloud. That is, the $8,393$ users that shared data in our trace are organized into $1,053$ distinct isolated connected components —corresponding to different companies and organizations. As can be observed in Fig. 5a, not all the connected components are equal in size; there is a large component consisting of 778 users, whereas the remaining ones are formed by 33 or less users.

In this sense, it is interesting to know if the largest connected component shares similar network properties with the rest of smaller ones. If we focus on *network clustering*, the answer to this question is negative; the left scatter plot of Fig. 5b shows that the network clustering coefficient of small connected components is generally much higher than the largest component. This indicates that groups of users belonging to *small organizations or reduced work groups tend to be more interconnected and collaborative* than large ones.

Similarly, the right scatter plot of Fig. 5b relates the network degree of users and their average *local clustering* coefficient. In Fig. 5b we clearly observe that users belonging to small connected components exhibit very high clustering coefficients, in particular for degrees ≤ 10. Apart of being less interconnected, Fig. 5b shows that the clustering coefficient in the largest component decreases as the node degree increases. This phenomenon is typical of *scale free networks* [21].

To check if the largest connected component is a scale free network, we need to inspect if its node degree distribution $P(k)$ follows a power law ($P(k) = k^{-\alpha}$) with $2 \leq \alpha \leq 3$ [21]. In practice, this yields that most users are connected with few other users, whereas a handful of them present high connectivity (i.e., *hubs*). As can be observed in Fig. 5c, the largest component node in-degree exhibits a linear down trend in the in log-log plot, which can be approximated by a power law distribution with $\alpha = 2.08$. However, for the out-degree distribution does not fit a power law with $2 \leq \alpha \leq 3$. Thus,



(a) Distribution of the number of files related to a sharing link (i.e., link strength).

(b) Relationship between the strength of user links and their degree.

(c) Network clustering of user communities sharing a single file.

Fig. 7. Analysis of user sharing links and the communities interested in the same file.

despite that the largest component of the NEC sharing network is not purely a scale free network, we can confirm that it *presents scale free characteristics*.

Insight 6: *The NEC sharing network is formed by many isolated components that represent different organizations and work groups. Some of them present totally different topologies.*

Insight 7: *There are sharing communities that present scale-free characteristics (e.g., power-law in-degree distribution).*

B. Identifying Content Producers and Consumers

Next, we aim at understanding the sources of shared content and how this content is disseminated within the NEC data sharing network. This can allow us to predict and anticipate data sharing flows within the system.

First, we want to analyze the strength of data sharing links across users. In particular, we measure link strength as the number of files that two users (or a user with himself) exchanged. To this end, Fig. 7a shows the strength of user sharing links depending on their type. As can be observed in Fig. 7a, whereas $\approx 99.8\%$ of sharing links consist of ≤ 10 files (i.e., weak links), a small fraction of them can be significantly stronger —up to 64 and 51 files for *in/out* and *self* links, respectively. These strong links may be object of optimization within the datacenter; for example, by co-locating users that share many files across them within the same servers [22].

As highly intense links cause a large fraction of the data sharing activity, an interesting question in this regard may be: do highly connected nodes exhibit strong links? Perhaps surprisingly, Fig. 7b illustrates that there is no correlation between the connectivity of users and the strength of their sharing links. Similarly, we have not found correlations between the degree of users and the number of stored files, as the resulting Pearson correlation coefficient is 0.1.

We also want to specifically inspect the activity of hubs in order to infer whether they are producers or consumers of shared content. In this sense, the top 0.1% of hubs are in general *producers of shared content*; in average, these users shared 46.3 files with others, whereas they requested 25.4 shared files. However, this behavior is rapidly diluted as we consider a larger number nodes; for the top 1% of highly connected nodes, the average number of produced and consumed shared files is 11.61 and 17.27, respectively.

Insight 8: *The strength of sharing links is skewed and hub users do not tend to present highly strong sharing links.*

Insight 9: *In NEC, main hubs are generally producers of shared content.*

C. Communities of Shared Content

Next, we analyze the size and clustering of groups of users sharing or interested in a given file. For the sake of clarity, in Fig. 7c we differentiated the content communities belonging to the large connected component from the rest of small components of the NEC sharing network.

Fig. 7c illustrates that the clustering of communities related to the same content presents a *more extreme behavior than the clustering of connected components* (see Fig. 5b). For the largest connected component, 94% of sharing groups related to the same file present a degree of clustering of 0. This means that most sharing interactions are done across small groups that have no common interest —master-slave sharing related to network hubs. Only a handful of shared content communities in the largest component present links related to other contents, which is denoted by a higher network clustering. Such a result suggests that this component may belong to a *big and hierarchical organization with independent departments*.

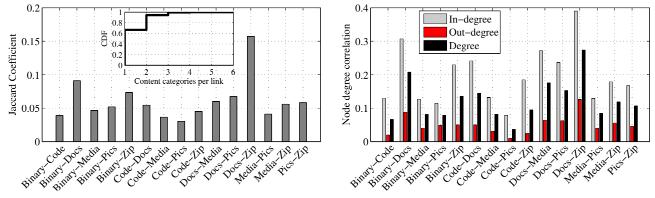
In the case of the smaller network components the situation differs; a majority of users sharing a file present very high clustering, yielding that they have other common sharing links (i.e., files). Also, many common interests of these user communities tend to be related to editable files (e.g., documents). This suggests that *peer-to-peer synchronization* techniques can be deployed in these small and clustered groups to offload the datacenter from data management relate to file syncing. This is specially true given that users in many small organizations coexists within high speed networks (e.g., LAN).

All in all, the fact of finding moderate/large groups of users (e.g., > 5) sharing the same file suggests that NEC might effectively exploit client-based content distribution [18], [20].

Insight 10: *The degree of clustering and the size of file-based communities greatly varies across network components.*

VI. Multi-layer Analysis of Link Types

Next, we resort to basic multi-layer network analysis to understand the data sharing network of NEC depending on different types of sharing links. This analysis will shed light on how to exploit data sharing interactions of users based on the content shared. Concretely, we define a multi-layer network in which each layer is formed by sharing links of a given file type (e.g., Docs, Media). Our focus is to look for similar topological characteristics across these layers (see Fig. 8).



(a) Link overlap between content-based network layers.

(b) Degree overlap between content-based network layers.

Fig. 8. Pairwise network overlaps indicating the similarity or dependence between interactions.

A. Link Overlap Between Layers

First, we focus on the link overlap across content-based layers. That is, we want to quantify the number of common sharing links across users that exist for more than one type of content. To this end, Fig. 8a shows the Jaccard coefficient of link overlap per pairs of network layers.

The most interesting point of Fig. 8a is the *Docs* category. Naturally, as documents are the most popular type of files in NEC, it tends to exhibit the higher amount of intersections with other file types, inducing a higher Jaccard coefficient. However, such a relationship is not lineal. To wit, the Jaccard value of the *Docs-Zip* tuple is significantly higher than the rest of tuples. This has a clear implication: *NEC users that share documents also tend to share compressed files*. We can also appreciate a similar correlation of with the *Binary-Docs* tuple.

On the other hand, we can deduce that most sharing links are generally not shared between tuples of content-based layers, as many Jaccard values fall below 0.05. This occurs even considering popular file types with a potentially number of links (e.g., *Pics*). One of the reasons for this phenomenon may be the *weak link strength*; that is, most sharing links are related with very few files (see Fig. 7a). Also, if we consider that 63.3% of links are related to a *single type of content* (see inner plot in Fig. 8a), this explains the low rate of common sharing links across layers. Thus, sharing links across users tend to be weak and specialized in terms of contents.

Insight 11: *There are types of contents that NEC users tend to simultaneously share with other users (e.g., Docs and Zip).*

Insight 12: *Most sharing links are weak and content-specific.*

B. Degree Correlation Between Layers

In Fig. 8b we inspect the Pearson correlation coefficient of node degree across multiple content-based network layers. The idea is to find whether nodes exhibit a similar number of sharing links depending on the type of shared files.

Appreciably, Fig. 8b shows that the degree of nodes is not correlated across network layers. That is, most correlation values fall below 0.3 that indicates no correlation at all. This seems a natural consequence of the specialization of link types; if most sharing links belong to a single type of content, the degree of nodes would be mostly concentrated on particular content-based layers. Therefore, the *sharing interactions of users are also closely related to specific types of files*.

As in the case of link overlap, the highest degree correlation values belong to the *Docs-Zip* tuple, followed by the *Binary-Docs* tuple. Moreover, we find that the in-degree of nodes exhibits higher correlation that the out-degree. The reason is that most users are content consumers, being much easier to find *in* links across network layers that in the case of *out* links.

Insight 13: *NEC users tend to share specific types of content.*

VII. Discussion and Lessons Learned

In this section, we discuss potential applications and benefits that our study may have on the NEC Personal Cloud, which may be of independent interest for other similar systems. Concretely, our objective is to propose optimized data management techniques (client and server sides) that exploit the social information of user data sharing interactions:

Clients can help to distribute content: The traditional Personal Cloud design presents a purely centralized architecture in which desktop clients communicate with the server for transferring files. However, the bandwidth costs for providers of such a centralized model is non-negligible [8], [23]. To reduce bandwidth costs, researchers may explore client-side strategies such as cooperative caching [24] and peer-to-peer content distribution [18] for Personal Clouds. In our view, the fat-client architecture of Personal Cloud desktop clients favors the deployment of this kind of strategies.

To inform this argument, we discovered that a small fraction of shared files and folders present a large amount of consumer nodes and normally very few contributors (*Insight 5*). Further, we found that shared files are large (*Insight 2*) and that highly connected hubs tend to produce shared content (*Insight 9*). These factors suggest that techniques like *cooperative caching* and *peer-to-peer content distribution* across consumers of shared data could greatly reduce bandwidth costs in Personal Clouds. Despite cooperative caching has been introduced in Web applications and Online Social Networks [20], the literature lacks from an exploration of these techniques in the Personal Cloud arena. Similarly, the incipient efforts on peer-to-peer content distribution for Personal Clouds can now benefit from exploiting social topologies of real users [18].

Towards distributed file syncing: File synchronization represents a large fraction of a Personal Cloud back-end activity [1], [2]. In most services, the communication overhead between clients and the server becomes specially heavy under frequent updates or when clients synchronize entire folders containing multiple small files [6]. Thus, we believe that such activity may also be optimized via peer-to-peer techniques.

In this sense, we found that documents —susceptible to frequent updates— represent a large fraction of the total number of files managed and shared in NEC (*Insight 1*). Moreover, many user communities are very clustered (*Insight 6*) —even in a file basis (*Insight 10*)— and they potentially coexist within high speed networks (e.g., LAN). These conditions are suitable to apply advanced *peer-to-peer syncing algorithms* in desktop clients. Today, only Dropbox integrates simple LAN-based syncing across a user's desktop clients[7]. However, more research is needed to bring this kind of synchronization to the next level; to wit, early attempts resorted to Bittorrent Sync[8] for synchronizing files in Personal Clouds [25]. In the

[7] https://www.dropbox.com/en/help/137

[8] http://www.getsync.com/

future, such kind of approaches might benefit from social and behavioral information of users to intelligently apply the most suitable syncing mechanism for a given file or folder.

Topology-aware data placement: In the context of Online Social Networks, remarkable research efforts advocated for merging the social structure of users with the layout or placement of data objects in the back-end [22]. In particular, SPAR [22] is a middleware that transparently partitions and replicates data objects based on the social graph structure of users to achieve data locality while minimizing replication.

In this sense, in our study we have shown that the social organization of NEC users is complex, leading to disparate data sharing topologies that may be handled differently. That is, there are groups of users organized in small and very interconnected networks, whereas other users form networks that present scale-free characteristics (*Insights 6 and 7*). Moreover, we realized that data sharing links across users may be very disparate in terms of intensity (*Insight 8*) and types of contents shared (*Insight 12*). We believe that the available social information of sharing interactions makes Personal Clouds a suitable target to apply *topology-aware data placement* strategies [22]. This may help to optimize the storage back-end and to reduce the storage costs of providers.

VIII. Conclusions

Data sharing accounts for an important fraction of the data management activity in a Personal Cloud. Motivated by the lack of studies, in this paper we contribute the first analysis of data sharing in a private Personal Cloud. In particular, we analyze sharing information of users gathered at the metadata back-end of NEC, an enterprise-oriented Personal Cloud.

First, we contributed a comprehensive study of the storage layer and data sharing network of NEC users. Our analysis provides novel insights regarding the storage characteristics of files in NEC compared with UbuntuOne, the activity of users, the concurrent usage of files and folders, as well as the different topologies that groups of users form based on their sharing interactions. Even more, we explored the similarities of sharing links of users based on the types of contents shared.

Thanks to our study of NEC, we proposed several data management optimizations that can be applied to similar systems by exploiting the storage and sharing interactions of users (e.g., cooperative caching, peer-to-peer syncing, etc.).

We conclude that this study and the released dataset may help on paving the way for a better characterization of data sharing in Personal Clouds. This, in turn, may provide a solid ground for researchers to devise novel data management techniques to further optimize these systems.

Acknowledgment

This work has been partly funded by the EU projects FP7 CloudSpaces (317555) and H2020 IOStack (644182) and Spanish research project "Cloud Services and Community Clouds" (TIN2013-47245-C2-2-R) funded by the Ministry of Science and Innovation. We also thank Albert Solé-Ribalta and Marc Sánchez-Artigas for their valuable feedback on the early stages of this paper.

References

[1] I. Drago, M. Mellia, M. M. Munafo, A. Sperotto, R. Sadre, and A. Pras, "Inside dropbox: understanding personal cloud storage services," in *ACM IMC'12*, 2012, pp. 481–494.

[2] R. Gracia-Tinedo, Y. Tian, J. Sampé, H. Harkous, J. Lenton, P. García-López, M. Sánchez-Artigas, and M. Vukolic, "Dissecting ubuntuone: Autopsy of a global-scale personal cloud back-end," in *ACM IMC'15*, 2015, pp. 155–168.

[3] T. Mager, E. Biersack, and P. Michiardi, "A measurement study of the wuala on-line storage service," in *IEEE P2P'12*, 2012, pp. 237–248.

[4] R. Gracia-Tinedo, M. Sanchez Artigas, A. Moreno-Martinez, C. Cotes, and P. Garcia Lopez, "Actively measuring personal cloud storage," in *IEEE CLOUD'13*, 2013, pp. 301–308.

[5] I. Drago, E. Bocchi, M. Mellia, H. Slatman, and A. Pras, "Benchmarking personal cloud storage," in *ACM IMC'13*, 2013, pp. 205–212.

[6] Z. Li, C. Wilson, Z. Jiang, Y. Liu, B. Y. Zhao, C. Jin, Z.-L. Zhang, and Y. Dai, "Efficient batched synchronization in dropbox-like cloud storage services," in *ACM Middleware'13*, 2013, pp. 307–327.

[7] Z. Li, C. Jin, T. Xu, C. Wilson, Y. Liu, L. Cheng, Y. Liu, Y. Dai, and Z.-L. Zhang, "Towards network-level efficiency for cloud storage services," in *ACM IMC'14*, 2014, pp. 115–128.

[8] J. Silber, "Shutting down Ubuntu One file services," http://blog.canonical.com/2014/04/02/shutting- down-ubuntu-one-file-services/, April 2014.

[9] G. Goncalves, I. Drago, A. P. Couto da Silva, A. Borges Vieira, and J. M. Almeida, "Analyzing the impact of dropbox content sharing on an academic network," in *XXXIII Brazilian Symposium on Computer Networks and Distributed Systems (SBRC)*, 2015, pp. 100–109.

[10] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *ACM IMC'07*, 2007, pp. 29–42.

[11] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *ACM IMC'07*, 2007, pp. 1–14.

[12] A. Iamnitchi, M. Ripeanu, and I. Foster, "Small-world file-sharing communities," in *IEEE INFOCOM'04*, vol. 2, 2004, pp. 952–963.

[13] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[14] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, no. 3, pp. 203–271, 2014.

[15] K.-M. Lee, B. Min, and K.-I. Goh, "Towards real-world complexity: an introduction to multiplex networks," *The European Physical Journal B*, vol. 88, no. 2, pp. 1–20, 2015.

[16] P. G. Lopez, M. Sanchez-Artigas, S. Toda, C. Cotes, and J. Lenton, "Stacksync: Bringing elasticity to dropbox-like file synchronization," in *ACM Middleware'14*, 2014, pp. 49–60.

[17] E. J. Whitehead Jr and M. Wiggins, "Webdav: Ieft standard for collaborative authoring on the web," *IEEE Internet Computing*, vol. 2, no. 5, pp. 34–40, 1998.

[18] R. Chaabouni, M. Sánchez-Artigas, and P. Garcia-Lopez, "Reducing costs in the personal cloud: Is bittorrent a better bet?" in *IEEE P2P'14*, 2014, pp. 1–10.

[19] R. Gracia-Tinedo, D. Harnik, D. Naor, D. Sotnikov, S. Toledo, and A. Zuck, "Sdgen: mimicking datasets for content generation in storage benchmarks," in *USENIX FAST'15*, 2015, pp. 317–330.

[20] S. Nikolaou, R. Van Renesse, and N. Schiper, "Cooperative client caching strategies for social and web applications," *LADIS'13*, 2013.

[21] X. F. Wang and G. Chen, "Complex networks: small-world, scale-free and beyond," *IEEE Circuits and Systems Magazine*, vol. 3, no. 1, pp. 6–20, 2003.

[22] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The little engine (s) that could: scaling online social networks," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 375–386, 2011.

[23] R. Gracia-Tinedo, M. Sánchez-Artigas, and P. García-López, "Cloud-as-a-Gift: Effectively exploiting personal cloud free accounts via REST APIs," in *IEEE CLOUD'13*, 2013, pp. 621–628.

[24] L. Ramaswamy, L. Liu, and A. Iyengar, "Cache clouds: Cooperative caching of dynamic documents in edge networks," in *IEEE ICDCS'05*, 2005, pp. 229–238.

[25] A. Lareida, T. Bocek, S. Golaszewski, C. Luthold, and M. Weber, "Box2box: A p2p-based file-sharing and synchronization application," in *IEEE P2P'13*, 2013, pp. 1–2.