On Pricing Online Data Backup

Laszlo Toka*† and Gergely Biczók†
*MTA-BME Information Systems Research Group
†MTA-BME Future Internet Research Group
Budapest University of Technology and Economics

***Abstract*—Online data backup services provide durable storage for user data without the hassle of traditional backup solutions, but comfort comes at a price. In this paper we focus on the costs of online data backup. First, we survey existing online backup pricing models: we compare unlimited and pro rata data plans, then we propose a novel, risk-based pricing scheme which suits risk-conscious users. Second, we build a simple Backup Selection Game where users can choose among these three cloud-based pricing schemes and the cost imposed by participating in a peer-to-peer backup system. We show that rational users generally prefer the pro rata scheme, irrespective of the characteristics of individual users. On the other hand, heterogeneous equilibria may also emerge under certain distributions of data volume per user, due to the effect of buying storage in bulk. We also discuss how peer-to-peer backup could emerge under a more elaborate cost model.**

I. Introduction

Backing up data is a vital, but potentially tedious process for both the everyday user and the large enterprise. Especially for home systems, most backup solutions require some technical savvy which can be prohibitive for the user. Luckily, virtualization technologies and plummeting hardware costs have enabled easy-to-use and high availability cloud storage services. In fact, services like Dropbox, Google Drive, Amazon S3 and Microsoft OneDrive are extremely popular and immensely successful in sharing and synchronizing data among devices, taking advantage of pervasive Internet connectivity.

Despite their success, cloud storage solutions have also demonstrated some shortcomings: data loss owing to inter-dependent failures [20], security mishaps due to configuration errors [26], or potential data theft [4]. One can argue that usability issues have now been solved, and harmful events are mostly mitigated by the evolution of cloud technology and emerging best-practices in managing these systems; however, there is one factor that never really goes away, namely the cost of storing large amounts of data over long periods of time. While the cost of storage per gigabyte is decreasing rapidly, the amount of data to be stored, analyzed and transmitted is doubling every two years. Big data means storage costs are likely to become and stay a significant component of IT budgets in the coming years.

Indeed, economic factors may lead to storage services shutting down in the future, similarly to what has happened to Drop.io [15]. In order to counter such events, and as an alternative to centralized online storage, peer-to-peer (P2P) and hybrid backup systems have been proposed [23]. While such systems may not be optimal for generic, remote file-system like operation, they are a good fit for data backup. In data backup, focus is shifted from data availability to data durability, i.e., guaranteeing that data are not lost. This comes with less stringent requirements in regard to access restrictions, adding redundancy, and since data is read only during a restore operation triggered by a disk failure, looser time limits. While P2P backup services are architecturally different, we argue that potential users are only interested in getting value for their money. This leaves cost as the deciding factor for choosing a given backup service.

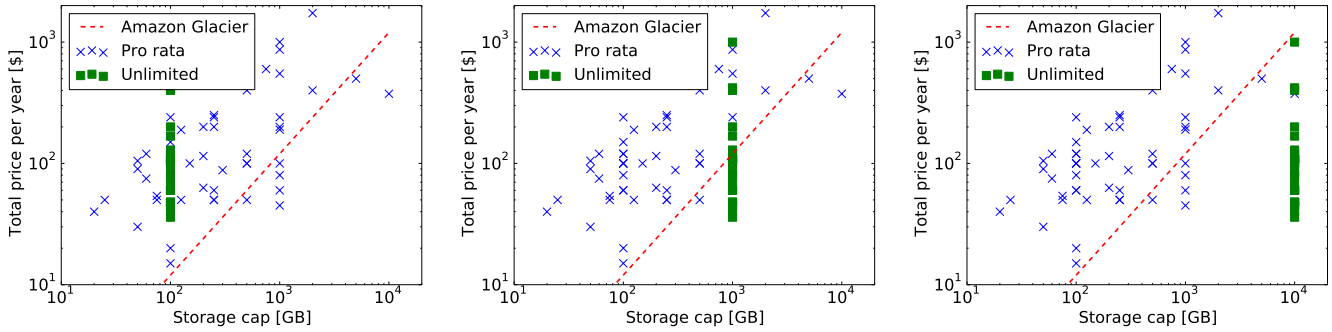
In this paper we investigate pricing and competition in the online data backup market. Our contribution is twofold. First, we survey the pricing methods of existing online backup services. We observe the proliferation of unlimited data plans, and empirically approximate the average data volume per user as expected by the corresponding service providers. Moreover, we propose a novel, risk-based pricing scheme. Second, we build a simple Backup Selection Game, where users can choose among three cloud-based and one P2P pricing schemes. We show that all users choosing the *pro rata* scheme is a favorable equilibrium irrespective of the characteristics of individual users. However, heterogeneous equilibria may also emerge under certain distributions of data volume per user, due to the effect of buying storage in bulk. We discuss how P2P backup could emerge under a more elaborate cost model.

The rest of the paper is structured as follows. Section II surveys the market for online backup services and establishes empirical inputs to our pricing model. Section III builds a simple game-theoretic model and analyzes its potential equilibria. Section IV discusses the limitations of our model and outlines potential future work. Finally, Section V concludes the paper.

II. Online data backup: technology and pricing

An online backup service provides users with an Internet-based system for backing up their files regularly (usually incrementally) and for restoring them in case of a hard disk failure. The goal of an online data backup system is to provide long-term and reliable (i.e., durable) storage for user data. This means they guarantee that data are not lost, but do not take the responsibility for availability, folder synchronization and such, which are characteristics of remote file-system like services such as Dropbox or Google Drive. Simply put, data from the provider are only read when a restore command is sent, implying a disk failure at the customer's premises. Adding to this, the restore process is not time-critical, especially that it requires the transfer of potentially huge amounts of data.

The work of Laszlo Toka was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (MTA).



(a) Unlimited estimated with cap at 100 GB (b) Unlimited estimated with cap at 1 TB (c) Unlimited estimated with cap at 10 TB

Fig. 1. Online backup services: prices vs. storage caps

A. *Cloud backup: existing vs. potential pricing schemes*

The most prominent of today's services use some form of cloud storage, providing an easy-to-use, dependable and reasonably priced solution. These services are usually accessed via a client software, which needs some configuration during installation, but runs automatically later on. This client software takes care of compression, encryption and data transfer. There are many service offerings on the market, each of them with a slightly different feature set based on the same core functionality. There are companies targeting specific market segments (enterprise or home users), but a lot of them serve all kinds of users differentiating via service levels. As it is common with online markets, backup providers usually offer a free/trial version of their services with limited storage space or no customer support to guide the users towards premium but paid versions or integrated targeted advertising. We ignore these alternatives and concentrate on companies which actually make money on providing a backup service. Here, we present a brief market survey [1].

Pro rata and tiered pricing. There are two online giants who offer clear usage-based (pro rata) pricing. Amazon has priced its Glacier backup service at $0.12 per gigabyte per year [5]. Google offers its Durable Reduced Availability (DRA) storage for double the price at $0.24 per gigabyte per year [16]. Smaller providers usually offer different tiers instead, e.g., Acronis, also a player on the home backup software market offers 1 TB for $189.99 per year ($0.19 per GB per year), 500 GB for $99.99 per year and 250 GB for $49.99 per year ($0.2 per GB per year); one caveat is that you also have to buy the client software for $49.99 [3].

Unlimited. A large number of companies offer unlimited plans, usually with some technical restrictions. These include policies on the number of computers, external hard drives and fair usage. One of the more popular providers, BackBlaze offers unlimited durable data backup (guaranteed for 50 years) for $5 per month ($60 per year), $50 per year or $95 per 2-year period ($47.5 per year) [7]. This plan is limited to one computer, although external hard drives connected to the same computer are allowed. Another company, CrashPlan offers a 2-year unlimited plan restricted to a single computer for $114.99 ($57.49 per year) and a 10-computer subscription of the same length for $289.99 (coming down to $14.99 per computer per year) [14]. One of the more interesting policies comes from MyPCBackup [19], who warns customers subscribed to their unlimited plan that the company may request them to transfer from the unlimited plan, if the company believes they are a business or that they have *exceeded the company's fair usage policy*. They also state that it is under MyPCBackup's sole discretion to judge that situation.

How much is "unlimited" anyway? Since backup providers have to buy or rent storage capacity to satisfy customer demand, they estimate the average "unlimited" data volume per user in order to operate efficiently and profitably. This raises the question: how much is "unlimited"? In order to get an answer, we have collected price quotes for 74 different data plans including 24 with unlimited data volume. Next, we plot their yearly prices against their storage caps, while assuming provider-estimated data caps for unlimited plans.

Results are depicted in Figure 1 with three different estimates for unlimited. We also plot the pure pro rata prices of Amazon Glacier, which is the cheapest solution also likely to be used by resellers. Note that, the bulk of the non-unlimited data points are to the left from the 1 TB line. More convincingly, in Figure 1(b), the relation between the Amazon Glacier price curve and the frontier for unlimited plans seems like the closest fit, taking into account the potential reselling of Glacier resources towards customers. From these figures we could guess that providers estimate the average "unlimited user" to have a data cap slightly below 1 TB. Note that this toy estimation assumes cost-based pricing with zero profit margin, and ignores costs other than for storage.

A post-pay, risk-based pricing scheme. Both unlimited and pro rata are insurance-type pricing schemes: users pay in advance (whether in annual or monthly installments) acknowledging and mitigating the risk of data loss. Providers can then calculate their prices based on the estimated data volume they have to store and potentially restore. We could reverse this thought process; what if the risk of data loss is deemed so low by users that they are willing to pay for it fully, but only when it actually happens? There is a twist in this operation mode in the case of data backup: no matter how and when the user pays, actual backups has to be made in order for lost data to be recovered some time in the future. Given an analogy to buying/not buying full-coverage car insurance, the difference is that the data backup company and the user have to sign a contract, and the company should start providing the service *before* the failure happens to enable future restoration.

Now, we let the providers calculate the risk of data loss

per user, and give them a price quote accordingly; this is the price users have to pay in the event of an *actual disk failure* at their premises. Obviously, providers have to somehow measure the risk: this could be based on industry standard disk drive failure rate metrics. There are several metrics around including the best-known Mean Time Between Failures (MTBF), a statistical term expressing the service hours between failures. Due to some issues with this metric, the hard disk industry has switched to Annual Failure Rate (AFR) [17], which specifies the probable percent of failures per year; AFR is based on the actually installed base of the given hard disk type. Furthermore, AFR presents providers with an easy-to-use metric when it comes to service plans. In order to get actual numbers, we turn to an in-depth study published by BackBlaze [18]. The study shows that drives exhibit an AFR of 1% to 10% in general with significant variations across brands but without a clear trend tied to capacity. The takeaway is that the price the provider should charge an unlucky user for the restoration of roughly one disk of data has to cover the storage costs of data on ca. 100 user disks per year.

B. P2P backup

P2P backup services have appeared in many forms recently; some of them are still operational. Here we give a short explanation on how P2P backup works, present the most important ones of such systems, and estimate their costs.

Technology. In a P2P backup system a user's data is always encrypted, split up to small chunks and redundancy encoded before leaving the local hardware [23]. Then the encrypted, redundant backup chunks are scattered on disks of fellow subscribers around the world. Each user has to add additional local storage capacity to their local hardware to store the backup data of other users also accounting for churn (users going offline/online), a characteristic of P2P systems. Apart from the cost of purchasing (and operating) these storage disk(s) there should be no extra cost to the service.

Existing solutions. Wuala [25] was the first commercial solution that allowed its users to offer storage capacity on their disks instead of paying for data backup. Wuala monitored the online availability of its users, checked their uplink and downlink capacities, and determined the volume of data the user could back up according to the offered storage space. Spacemonkey [21], a Kickstarter project in 2013, ships an external storage device to its subscribers providing the user with 1 TB of storage; afterwards the P2P service costs $10 per month or $49 per year. Following the same concept, Connected Data [13] offers dedicated devices called Transporter and Transporter Sync to users who want to join their P2P backup network for $159 and $99. Neither of those include any hard drives. Symform [22], an otherwise regular online backup service provider, offers, similarly to Wuala, the possibility of contributing storage space instead of paying for cloud storage. If the user allocates 2 TB of storage on its device, it receives 1 TB of storage for data backup at the provider. Finally the last group of P2P backup solutions include those that give away the backup software, with the user being responsible for providing the hardware. Both CrashPlan [14] and BuddyBackup [9] make their application freely available, which periodically detects changes in files, incrementally encodes them, encrypts the data, and uploads pieces of it to a predefined list of hosts over the Internet.

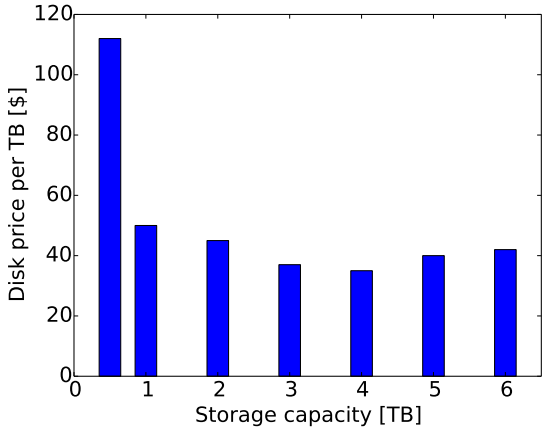


Fig. 2. Disk prices in November 2014

Cost. All P2P backup solutions incur storage device costs, hence we collected the lowest prices for disks of various capacity at BestBuy [8], a major US retailer. As it is shown in Figure 2, in 2014 the end customer can buy storage disks for almost as low as $30 per TB. Note that we have picked the cheapest hardware, although we did not see much difference between the prices of disks of form factor 2.5" and 3.5", or internal and external disks. Assuming a negligible cost of power and a disk amortization period of 5 years, the cost of having 1 GB of storage space is ¢0.6 per year or ¢0.05 per month. Note that tape prices are even lower (e.g., an LTO Ultrium 6 cartridge holds 2.5 TB and sells for about $50), but they require a costly apparatus, hence we ignore them. By taking the price of 1 GB of durable storage at Amazon Glacier as reference, it is a bit of a surprise why P2P data backup has not gained significant momentum so far: ¢1 vs. ¢0.05 per month is a 20-fold difference. In Section III we try to find an answer to this question from an economic perspective.

III. The Backup Selection Game (BSG)

After our market survey, we are now ready to build our pricing model and game.

A. Assumptions

We have made several simplifications to construct a tractable model. We assume cost-based pricing with a zero profit margin, and account only for storage costs. We model the pricing schemes as different providers, and we assume that each provider is capable of serving the whole market. We ignore free data plans and per computer pricing, and we view tiered plans as special cases of pro rata, hence do not analyze it as a separate scheme. We ignore compression and redundancy in data storage, as different schemes will likely employ similar measures. We do not cover the temporal evolution of the online backup market, and opt for a single shot game; hence we also ignore providers having downtimes or going out of business. Finally, we model the users as rational decision-makers, and do not cover behavioral aspects. The implications of some of these assumptions are presented in Section IV.

Unit price of storage. We build function $p(x)$, the estimated yearly cost of storing data volume x, on Amazon Glacier's pro rata price and fit a logarithmic curve to the decay of Amazon S3 [6] prices vs. storage demand. We also account for the actual disk prices that are 20 times cheaper than Glacier prices (assuming 5-year amortization). This way $p(x)$ yields ¢0.6 as the annual unit cost of 1 GB storage, ¢0.588/GB for 1 TB and ¢0.576/GB for 1000 TB. Function $p(x)$ is given as follows: $p(x) = 12 \cdot (0.03 - 2 \cdot 10^4 \log(x))/(3 \cdot 20)$.

Disk failure rate. We define the disk failure function $r(x)$ to be as simple as possible. We assume an AFR of 1% and the usage of 4 TB user disks at user premises as they represent both the cheapest (per capacity unit) and most reliable storage alternative today. We argue that users with more data pose a larger risk owing to more/larger disks. Then the probability of more then zero failed disks out of n, by neglecting further elements of the binomial formula, is $\mathbb{P}(i > 0) = 1 - \mathbb{P}(i = 0) = 1 - (1-p)^n \approx np$. For a given user, n is calculated out of x by assuming 4 TB disks, so $r(x)$ is given as follows: $r(x) = 0.01 \cdot x/(4 \cdot 1024)$.

B. Players, strategies, cost functions

The players of BSG are users who want to use online backup services. Users are characterized by their backup data volume. Strategies of BSG simply consist of choosing from four available pricing schemes (and hence four "providers"): *unlimited, pro rata, post-pay and P2P*. For each pricing scheme we define a cost function. In the case of cloud backup strategies, i.e., *unlimited*, *pro rata* and *post-pay*, we model the total cost of the service provider as proportional to the aggregate volume of backed up user data. The difference between these three schemes is how the total cost is dispersed among users. The *P2P* strategy does not involve a central provider, users only pay for their own hardware.

Unlimited. The *unlimited* scheme distributes the total cost equally, with the total cost being $p\left(\sum_{i\in U} x_i\right) \sum_{i\in U} x_i$, where x_i is the volume of backup data of user i, $p(x)$ is the non-increasing unit price function of storage for capacity x, and U is the set of users who choose the *unlimited* scheme. Therefore the cost of participation is:

$$c_i = p\left(\sum_{j\in U} x_j\right) \frac{\sum_{j\in U} x_j}{|U|} \quad \forall i \in U. \tag{1}$$

Pro rata. Those who choose the *pro rata* scheme, constituting the user set denoted by R, pay as they go:

$$c_i = p\left(\sum_{j\in R} x_j\right) x_i \quad \forall i \in R. \tag{2}$$

Post-pay. In the *post-pay* pricing scheme, only users, who are actually hit with a disk error and retrieve their backup from the provider, pay. In this case, the division of costs is based both on the volume of data and the failure rate $r(x)$, a non-decreasing function of the underlying volume of data. Hence the probabilistic cost of user i who belongs to this set O is:

$$c_i = \sum_{j\in O} x_j \cdot p\left(\sum_{j\in O} x_j\right) \frac{r(x_i)\, x_i}{\sum_{j\in O} r(x_j)\, x_j} \quad \forall i \in O. \tag{3}$$

P2P. Finally, the users of the P2P network, denoted by P, pay only for the disk they contribute:

$$c_i = p(x_i)\, x_i \quad \forall i \in P. \tag{4}$$

C. Homogeneous equilibria

Naturally, we are interested in the Nash equilibria of the game above. As expected from the structure of the cost (payoff) functions, the number of users choosing the same scheme will be a main factor in deciding the stable outcome. First, let us investigate the special cases, where all users are grouped together in the same scheme.

Proposition 1. *(Unlimited NE) The strategy profile of all users selecting the unlimited scheme is a (weak) Nash equilibrium iff* $\forall i \quad c_i = p\left(\sum_{j\in U} x_j\right) \frac{\sum_{j\in U} x_j}{|U|} \le p(x_i)x_i$.

Proof. (sketch) If a single user deviated from the strategy profile, its cost would be $p(x_i)x_i$ in either one of the other schemes. If the condition above holds, no users have the incentive to deviate. On the contrary, if the condition did not hold, at least one user would have the incentive to deviate unilaterally, since switching to either one of the three remaining strategies would decrease her cost. □

An example for this situation is when all users have the same volume of data: $\forall j \quad x_j = x_i$. In this case, all users pay $c_i = \frac{p\left(\sum_{j\in U} x_j\right)\sum_{j\in U} x_j}{|U|} = p\left(\sum_{j\in U} x_j\right) x_i < p(x_i)x_i$, since $p(x)$ is monotone decreasing. Note that if there are users with significantly less data than the average, those would switch to other schemes, avoiding the subsidization of heavy users.

Proposition 2. *(Post-pay NE) The strategy profile of all users selecting the post-pay scheme is a (weak) Nash equilibrium iff* $\forall i \quad c_i = \sum_{j\in O} x_j \cdot p\left(\sum_{j\in O} x_j\right) \frac{r(x_i)x_i}{\sum_{j\in O} r(x_j)x_j} \le p(x_i)x_i$.

Proof. The same as in Proposition 1. □

An example for this situation is the same as for Proposition 1. Generally, if the heaviest users have only slightly more data than average users, *post-pay* NE can emerge.

Theorem 1. *(Pro rata NE) For any data volume distribution* $(x_0, ..., x_n)$*, where* $\forall i \quad x_i > 0$ *and* n *is the number of users, the strategy profile of all users selecting the pro rata scheme is a (strong) Nash equilibrium.*

Proof. (sketch) In this case, any given user would pay $c_i = p\left(\sum_{j\in R} x_j\right) x_i$. If one of them deviated and chose either one of the remaining three schemes, her cost would change to ${c_i}^* = p(x_i)x_i$. Now, $c_i < {c_i}^*$ for any user i, since the function $p(x)$ is monotone decreasing. □

All users choosing the *pro rata* scheme equals to a proportionally fair allocation of costs, and is a NE for any sensible user population. Note that situations depicted in Proposition 1 and 2, are close-to-fair allocations. In the special case of all users having the same data volume, the *unlimited* and *post-pay* schemes converge to the *pro rata* scheme.

Theorem 2. *(P2P NE) For any data volume distribution* $(x_0, ..., x_n)$*, where* $\forall i \quad x_i > 0$ *and* n *is the number of users,*

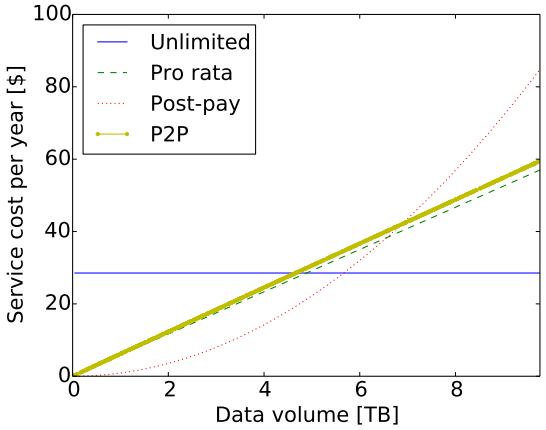


Fig. 3. Cost vs. data volume in homogeneous equilibria with uniform data volume distribution

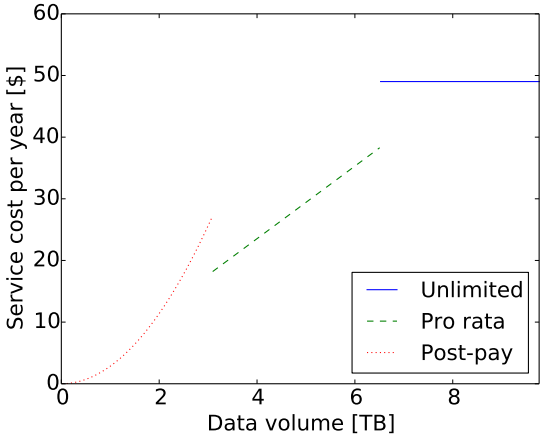


Fig. 4. Costs when users are equally divided among the central services along their backup volumes

the strategy profile of all users selecting the P2P scheme is a (weak) Nash equilibrium.

Proof. (sketch) Here, each user has the same cost $c_i = p(x_i)x_i$. If a single user wants to deviate from the strategy profile, her cost would be exactly the same under the *unlimited*, *pro rata* and *post-pay* strategies. □

We can notice that the individual costs and the total system cost in the *P2P* NE are higher than in the *pro rata* NE. In fact, for Nash equilibria presented in Proposition 1 and 2 and Theorem 1, the total equilibrium cost equals to the social optimum. This follows from the fact that the *unlimited*, *post-pay* and *pro rata* NE realize the same optimal total system cost among the users, capitalizing on the lowest possible unit storage price. On the other hand, the *P2P* scheme does not benefit from the economies of scale from buying storage capacity in bulk. Thus, we can readily characterize the Price of Anarchy (PoA, the fraction of total cost in social optimum vs. in the worst-case equilibrium) and the Price of Stability (PoS, the fraction of total cost in social optimum vs. in the best-case equilibrium).

Corollary 1. *(PoA and PoS) In the Backup Selection Game, the Price of Anarchy is* $\frac{p\left(\sum_i x_i\right)\sum_i x_i}{\sum_i p\left(x_i\right)x_i}$*, and the Price of Stability is* 1.

In Figure 3 we visualize the cost against the volume of backup data in all four pricing schemes, assuming that every user (1000 in this example) selects the same scheme (data volumes are uniformly distributed). It is clear that the *unlimited* scheme favors the heavy users, and the *post-pay* scheme the light users. The total income of the central provider schemes is always the same (the total storage cost), and the *P2P* scheme is always costlier then the *pro rata* scheme, due to the decreased storage unit price when purchasing in bulk.

D. Data volume distribution and heterogeneous equilibria

Simulation: different distributions of x. Here we study the effects of applying various distributions of the volume of backup data across users. We consider the value of x throughout the user set to be drawn from a uniform, normal or Pareto distribution. We set the minimum volume of backup to $\min = 10$ GB, the maximum to $\max = 10$ TB per user; if a randomly generated parameter falls outside this segment, we set it to the minimum or maximum, respectively. For the normal distribution, we set the mean to $\frac{\max+\min}{2}$ and the variance to $\frac{\max-\min}{6}$. For the Pareto distribution, we set the shape parameter to $\frac{\max+\min}{\max-\min}$, so the mean approaches those of the uniform and normal distribution.

We wonder if a heterogeneous Nash equilibrium across pricing schemes can emerge from the game. Our motivation is that light users seem to pay less with the *post-pay* scheme, heavy users usually prefer a flat-rate unlimited offering, while medium users could be best off with a usage-based (*pro rata*) scheme. We refer the reader to Figure 4: here $10,000$ users are distributed evenly across schemes and uniformly in data volume. Light users assigned to *post-pay*, medium users to *pro rata* and heavy users to *unlimited* schemes, respectively. Could there be an equilibrium similar to this setting?

Now, we run iterative simulations (1000 rounds) in which the users select their best response strategies sequentially in a heuristic, simulated annealing manner: there is a certain amount of randomness in which strategy they choose in each round, but it gradually diminishes by the end of the simulation. This is necessary in order to avoid artifacts of the initial selection of pricing schemes: at start each user is randomly assigned to given schemes with equal probability. We find that no matter what the volume distribution is, the final market will not be segmented: all data volume distributions (uniform, normal, Pareto) lead to the monopoly of the *pro rata* provider.

Distributions with tiers. Now, we turn to distributions of x, where there are clearly distinguishable tiers. We consider a scenario with 112 users, where 102 light users have data volume of $x_1 = 1$ TB and choose the *post-pay* scheme, while 10 heavy users have $x_2 = 10$ TB and choose the *pro rata* scheme. In this case light users will not deviate to either *pro rata* with ${c_1}^*$ or *unlimited* with ${c_1}^{**}$: $c_1 = \frac{102 \cdot p(102)}{102} = p(102) < {c_1}^* = p(101) < {c_1}^{**} = p(1)$. For heavy users, $c_2 = p(100) \cdot 10 < {c_2}^* = p(10) \cdot 10 < {c_2}^{**} = \frac{112 \cdot p(112) \cdot 100}{202} = p(112) \cdot 55.44$, thus they will not deviate either (${c_2}^{**}$ being the *post-pay* and ${c_2}^*$ the *unlimited* costs). Hence this allocation is a heterogeneous Nash

equilibrium with users allocated to two different strategies.

Conjecture 1. *(Heterogeneous NE) For any strategy profile* $(s_1, \ldots, s_n)$*, excluding the coexistence of P2P and pro rata, there exists an underlying distribution of user backup volumes* x *such as* $(s_1, \ldots, s_n)$ *is a Nash equilibrium.*

Note that the distribution of user backup volumes x has to satisfy specific requirements with regard to appropriately sized gaps in data volumes. We leave the more formal expression and proof of this conjecture to future work.

IV. Discussion

Here we list and discuss some limitations of our model.

Provider cost and revenue model. By concentrating on storage cost only, we have simplified the cost model of providers. In reality, these providers face several other types of cost owing to egress/ingress bandwidth, power, maintenance, emergency upgrades and unused capacity [2]. On the other hand, we have also assumed that revenues only cover costs, hence the profit of providers is zero. While these simplifications do not affect the competition among cloud providers, they do have a profound effect on the proliferation of P2P backup. We have seen that buying in bulk assures the superiority of *pro rata* over *P2P* in Section III-C. However, if we factor in other costs and a profit margin for central providers (represented by a coefficient α), it could be enough to tip the market towards P2P. More precisely, *P2P* dominates *pro rata* if $\forall i \quad c_i = \alpha \cdot p\left(\sum_{j \in R} x_i\right) x_i > p(x_i) x_i$. In fact, depending on the bulk capacity price decay, a coefficient as low as $\alpha = 2$ is sufficient for the market to tip with all uniform, normal and Pareto backup volume distributions.

Users like flat-rate. Several studies conducted in the context of Internet access showed that users prefer simple flat-rate pricing, e.g. [24]. It is also well-known, however, that Internet access providers are not satisfied with flat-rate pricing, since it has resulted in declining sales volumes, shrinking markets and a shift in competition from service quality to price [12]. Adding to this, the steep growth in fixed and especially mobile Internet traffic makes it even harder to maintain all-you-can-eat pricing. Transferring this line of thought to online data backup, it seems that backup providers are also facing the same problem. Cloud storage costs will never reach zero: adding a user represents a low marginal cost, but at a given point the provider has to make a sizable investment for expanding its infrastructure, resulting in a significant fixed cost [10]. This could make backup providers experiment with pro rata, value-based, risk-based or other smart pricing schemes, even though they might be slightly too complicated for the user to understand. Raising the price of unlimited plans is not a solution due to the competition in the market [1].

Realistic distribution of user data volumes. It is clear from our analysis that the distribution of data volume to be backed up dictates the evolution of the market. Distributions with observable tiers might emerge: storage media are available in only a few well-defined sizes and this shapes the backup demand of the user to a certain extent. Also, there are studies with regard to residential Internet traffic which imply the existence of a heavy-tailed, Pareto-like distribution for user traffic demand, see [11]. Extrapolating from these two observations, we argue that the real-world distribution could also exhibit a Pareto-like slope on a few actual values with non-negligible probability. Interestingly, such a distribution could potentially enable the coexistence of multiple pricing schemes in a market equilibrium. Nevertheless, obtaining a real-world dataset from an online backup provider would be the ultimate way to clear the remaining uncertainties.

V. Conclusion

In this paper we have studied some facets of the market for online data backup. First, we showed the proliferation of unlimited data plans and presented a cost-based approximation of how providers could perceive "unlimited". Second, we constructed an empirically rooted, simple Backup Selection Game, where rational users select their providers based strictly on price. We found that among the four competing pricing schemes – *unlimited, pro rata, post-pay and P2P* – users generally prefer *pro rata* independent of the underlying distribution of data volume per user. In special cases, equilibria with multiple pricing schemes may also emerge under certain distributions of individual data volume, due to the effect of buying storage in bulk. Although the *P2P* scheme does not fare well in our simple game, we discuss how it could emerge under a more elaborate cost model.

References

[1] 37 Online Backup Services Reviewed. About Technology. http://pcsupport.about.com/od/maintenance/tp/online_backup_services.htm.
[2] A framework for comparing CAPEX to OPEX storage alternatives. Ovum. https://assets1.csc.com/infrastructure_services/downloads/A_Senior_Executive_s_guide_to_Storage_as_a_Service.pdf.
[3] Acronis. http://www.acronis.com.
[4] Advanced Password Cracking–Insight. Elcomsoft Blog. http://blog.crackpassword.com/tag/icloud/ Last accessed: Nov 2014.
[5] Amazon Glacier. http://aws.amazon.com/glacier/.
[6] Amazon S3 pricing. http://aws.amazon.com/s3/pricing/.
[7] BackBlaze pricing. https://secure.backblaze.com/buy.htm.
[8] BestBuy. http://www.bestbuy.com.
[9] BuddyBackup. http://www.buddybackup.com.
[10] Can cloud storage costs fall to zero? Enterprise Storage Forum. http://www.enterprisestorageforum.com/storage-management/can-cloud-storage-costs-fall-to-zero-1.html.
[11] K. Cho, K. Fukuda, H. Esaki, and A. Kato. The impact and implications of the growth in residential user-to-user traffic. SIGCOMM. ACM, 2006.
[12] Cisco. Rethinking flat rate pricing for broadband services how service providers can monetize internet traffic growth via value-based pricing. *Cisco Whitepaper*, 2012.
[13] Connected Data Transporter. http://www.filetransporter.com.
[14] CrashPlan. http://www.code42.com/crashplan/.
[15] Drop.io shutdown. Drop.io blog. http://drop.io/cloud-storage/.
[16] Google Durable Reduced Availability Storage. https://cloud.google.com/storage/docs/durable-reduced-availability.
[17] Hard disk drive reliability and MTBF / AFR. Seagate. http://knowledge.seagate.com/articles/en_US/FAQ/174791en?language=en_US.
[18] Hard drive reliability update. BackBlaze. https://www.backblaze.com/blog/hard-drive-reliability-update-september-2014/.
[19] MyPCBackup. http://www.mypcbackup.com.
[20] Online backup company Carbonite loses customers' data, blames and sues suppliers. TechCrunch. http://tcrn.ch/dABxRn.
[21] Spacemonkey. https://www.spacemonkey.com.
[22] Symform. http://www.symform.com.
[23] L. Toka, M. Dell'Amico, and P. Michiardi. Online data backup: A peer-assisted approach. In *P2P*. IEEE, 2010.
[24] H. R. Varian. The demand for bandwidth: Evidence from the index project. *Broadband: should we regulate high-speed Internet access*, pages 39–56, 2002.
[25] Wuala. https://www.wuala.com.
[26] Yesterday's authentication bug. http://blog.dropbox.com/?p=821.