Diversity Decay in Opportunistic Content Sharing Systems

Liam McNamara
Computer Laboratory
University of Cambridge
liam.mcnamara@cl.cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge
salvatore.scellato@cl.cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cecilia.mascolo@cl.cam.ac.uk

***Abstract*—As content that users access on their mobile devices becomes bulkier, opportunistic networking is becoming a potential complement to centralised and infrastructure based downloads. We study how users share items of mutual interest with each other with a simple model based on a 'networked urn process'. We investigate the effect of different content sharing policies upon a multi-category set of items. We find that the process of sharing mutual interests inherently disproportionately reinforces category replication disparity, i.e., the most popular categories become proportionally even more numerous.**

These findings uncover a major hurdle in the creation of automatic opportunistic file sharing between users. Even if users altruistically sacrifice battery power and network resources to share content not relevant to them, overall, the system may not be able to fairly distribute items that belong to niche categories.

I. Introduction

Users accessing content (e.g., music or video) on their mobile devices are becoming more demanding about the content's length and quality. They expect the same user experience that they have on their desktops while watching videos, news clips and listening to music. Multimedia traffic puts the cellular infrastructure under heavy stress, badly affecting the perceived quality of service [1]. As a result, opportunistic networks (i.e., with ad hoc sharing) may be put in place to locally serve bulkier traffic [2].

Various approaches to model how content spreads in opportunistic networks have been proposed, often using the framework of epidemic spreading models [3]. It may seem natural to employ such epidemic modeling techniques when considering the spreading of data items, this does not consider bandwidth: the use of many superimposed epidemic spreading processes would imply infinite bandwidth, since the occurrence of one interaction would not impact whether another one is performed. We require a model that uses many interconnected users, able to share data items between each other while considering also their own preferences. Each user is interested in multiple categories from a population of data categories. They should be able to consecutively acquire items of these categories from their neighbouring elements.

Each sharing interaction involves sampling the categories of another source device and copying content from it. We represent this sampling by adopting an *urn process* [4]. Such a process is defined as an *urn* containing a set of *coloured balls*: then the balls can be drawn and the content of the urn manipulated to study a particular phenomenon of interest. For instance, consider an urn initially containing one red ball and one blue ball: a ball is randomly chosen from the urn at each time step and then it is replaced together with an additional ball of the same colour. Thus the number of balls keeps increasing and a 'rich-get-richer' feedback process takes place for the selected colour, resembling the effect of popular content becoming increasingly more popular [5]. Only a ball of a colour that both urns already possess can be duplicated, in order to mimic categories of interests. Furthermore, each ball is numbered, balls of matching colour and number are considered identical, representing different copies of the same piece of media. If a ball of a given colour and number is already possessed, it will not be copied.

The library of a device can only be populated from neighbouring devices, we will use a *time-varying networked urn process* metaphor, where the ball sampling happens across different urns connected by the edges of a time-varying graph. The graph dynamics are due to the fact that the set of available neighbours' devices that can be selected for sharing changes over time, since connectivity changes.

A. Sharing Policies

We modify a standard networked urn process with the following sharing policies. Note that these policies act on the intersection between the colours present in the source and in the target urn. Once a policy has chosen the category, a track that is not possessed is uniformly randomly selected.

The policies we investigate are *Random*, where the colour to share is chosen at random, *Popular* and *Unpopular*, where the colour is chosen with probability respectively directly and inversely proportional to the number of balls in the urn. We also study two policies which exploit full knowledge of the system to pick the most popular, or unpopular, colour: *Oracle Popular* and *Oracle Unpopular*.

II. The Sharing Process

This section gives a more precise description of the device behaviour model we are proposing.

A. Network of Urn Processes

We consider the system as a set of n urns $V = \{u_1, ..., u_n\}$, each initially containing a subset of coloured

978-1-4577-0351-5/11/$26.00

numbered balls. Each urn u_i has an initial number of different balls from a subset of colours $C_i \subseteq C$, with $g = |C|$, and will only wants balls from this set of colours. If an urn already contains a ball of a certain colour and number, another ball of matching colour and number will never be duplicated in the urnThe model evolves over time along sequential *ticks* with each urn trying to collect all numbers of each colour that it is interested in. Urns are connected in a graph structure which evolves over time and they can only duplicate balls from other connected urns.

At each time tick, all urns not engaged in sharing will choose a random *available* neighbour that shares a colour of interest, in order to initiate exchange. Each urn makes their source selection in a random order, to avoid synchronisations. An exchange is initiated and the source urn is set as *unavailable*: note that urns unavailable to act as a source will still search for others to download from. If no neighbours are available, the urn waits *idly*. The choice of the colour to be duplicated is made according to the selection policy that is being used, as was described in Section I-A. Whenever a ball exchange happens over a connection, it lasts for D ticks: if the connection disappears before the exchange is completed, the ball is not duplicated and the urn may select a new source. This feature mimics the actual downloading process in real systems, which would abort when connectivity is lost before completion. After D time ticks the ball of the chosen colour is duplicated in the target urn and a new exchange may take place, with the same source or with another.

In order to operate some of the sharing policies each urn records the popularity of different colours to inform their choices. Each urn stores this information in a queue of length h. In this structure, the last h unique urns that have interacted with the urn will have an entry detailing what their content was. These entries contain the proportion of various colours found in the different urns. This queue will then be aggregated to give a sampled approximation of the global colour popularity.

B. Altruistic Behaviour Modeling

We also model altruistic behaviour in our system: if an urn decides to perform an altruistic exchange, it will accept a ball without considering its colour. This behaviour allows unpopular colours to be replicated by urns that may not be interested in them. Then they will be able to act as an intermediary in the distribution of the colours that do not spread very well due to the pairwise incompatibility. Just because an urn is being altruistic does not mean it will always receive a colour it is not interested in: it just that the sharing policy will operate on the totality of the source's interests, rather than just on their intersection of tastes. To control the level of altruism we introduce a parameter p_A, which is the probability that a given ball exchange is altruistic. We will show how altruistic behaviour dramatically impacts the system's final content distribution.

Parameter	Symbol	Default Value
Download Duration	D	200
Global Categories	G	100
Selected Categories	g	5
Category Files	T	100
User Library Size	M	100
Zipf Exponent	a	1
Altruism Probability	p_A	0
History Size	h	10

Table I
DEFAULT VALUES FOR THE PARAMETERS OF THE MODEL.

III. EVALUATION

The results from the simulation of our model are presented in this section. The major parameters that control the behaviour of the model are defined in Table I together with their default values. The default time it takes to perform a download D is set to 200 seconds, equivalent to 5MB files transferred at 200Kbit/s. The Zipf-like category popularity distribution has an exponent $a = 1$. The evaluation of our model is performed on a real-world dataset of contact traces from Transport for London (TfL), described with greater detail in [6]. This dataset includes journey traces collected in the London Underground over one month containing two separate train lines at 2007's end, comprising of 59 Million of journeys made by 200,000 unique passengers.

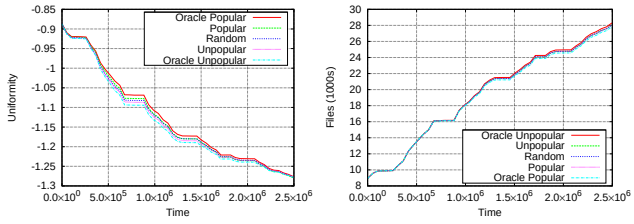
A. Performance Metrics

If there are a few categories that enjoy massive popularity, then the system is more polarised than when most categories have about the same share of files. It is useful to be able to distill the uniformity of the category popularity distribution into a scalar value. We define *Uniformity* as the estimated exponent of the category frequency distribution, computed as the gradient of a *linear least square regression* of its log-log plot. This gradient is akin to a power law's exponent: it gives a reasonable estimator of the skew of the popularity distribution. When uniformity is at its maximum value 0, the distribution of category items is completely uniform, and as it decreases into negative values the categories have greater disparity of popularity. Although each urn initially has an i.i.d. probability to choose its colours of interest, some order will still arise out of the initial random conditions. The most popular colour will then have its popularity reinforced through the feedback effect of the sharing process.

B. Simulation Results

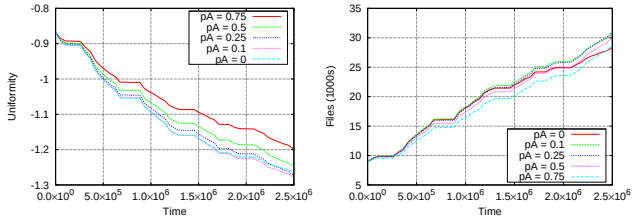
We now present the sharing process simulation results.

1) Uniformity Evolution: The performance of each sharing policy is shown in Figure 1. The most striking aspect of the system Uniformity (Figure 1(a)), is that *the uniformity of the system is always decreasing regardless of the adopted sharing policy.* Even the *Oracle Unpopular* policy, which will (with perfect global knowledge) share the least popular categories between two peers. The total amount of files in the system's libraries over time is plotted in Figure 1(b). There



(a) Library Uniformity (b) Library size

Figure 1. Change in uniformity (a) and library size (b) over time for different sharing policies.



(a) Library Uniformity (b) Library size

Figure 2. Change in uniformity (a) and library size (b) over time for different sharing policies with different degrees of altruistic behaviour.

are slightly more files gained by policies that favour unpopular selections. As when choosing between a popular and unpopular category, it is preferable to choose the unpopular category, which will be harder to find in future. Unpopular policies exploit more opportunities, because when a user's unpopular tastes are unavailable, they should still be able to collect the widespread popular categories.

2) Altruistic behaviour: After seeing that decreasing uniformity can not be stopped even when favouring unpopular file exchanges, the effect of including altruistic transfers is depicted in Figure 2. Each curve shows different altruistic probabilities. As expected, $p_A = 0$ are the same curves from Figure 1, as no altruistic transfers occur. In Figure 2(a) it can be observed that when using altruism there is a slower degradation of uniformity. Figure 2(b) shows that *with a small amount of altruism each device is on average able to gather more files*. Though very high altruism rates ($p_A = 0.75$) cause a reduction in library growth, due to altruistic transfers not being spent increasing useful files.

IV. Implications

Our results have shown the troubling effect of a tendency for homogenisation in automatic distributed sharing systems. We used a relatively straight-forward sharing model, where the more popular categories were replicated far in excess of less popular ones. An obvious approach to mitigate the extreme proliferation of popular tastes, by the preferential sharing of less popular items, does not reverse this process, only slows it. Going further and having nodes replicate unpopular files that they are not even interested in still does not reverse the trend. Though a small amount of altruism decreases diversity decay and usefully increases file distribution.

V. Related Work

Relevant research on this topic includes different approaches to information sharing on opportunistic networks and abstract models of content or culture propagation. Studies on content spreading in opportunistic networks have also been proposed, in [2] a model of the mean content update age is given. With respect to these works we take a different stance, assuming content is bulk downloaded instead of streamed or directed to specific users. Rather than assuming each user will manage their device to obtain specific content, we envisage an automatic system where devices only need to identify categories of interest.

A suitable formulation of different tastes spreading over a network has been considered in the field of cultural spreading. Axelrod's model can be further generalised as Networked Urn Processes [7], where nodes represented by an urn filled with multiple coloured balls. The authors show how social influence/selection interact in social processes in different online social scenarios, the aim of our model is to understand a data sharing system.

VI. Conclusions

This paper indicated there is a naturally emergent behaviour of library homogenisation, it may not be unavoidable, but will require some tough choices. An obvious approach is to avoid creating replications of popular categories, i.e., not downloading files, even when it is possible. This could considered contrary to the aims of a *dissemination* system. To conclude, automated content sharing may only be feasible for the most popular data items.

References

[1] New York Times, "iPhones Overload AT&T Network, Angering Customers," September 2009.

[2] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and Scalable Distribution of Content Updates over a Mobile Social Network," in *Proc. of INFOCOM'09*, April 2009.

[3] W. O. Kermack, A. G. McKendrick, and P. McKinlay, "Death-Rates in Great Britain and Sweden," *Journal of Hygiene*, vol. 34, no. 4, pp. 433–457, 1934.

[4] R. Pemantle, "A survey of random processes with reinforcement," *Probability Surveys*, Feb 2007. [Online]. Available: http://arxiv.org/abs/math/0610076

[5] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.

[6] L. McNamara, C. Mascolo, and L. Capra, "Media Sharing based on Colocation Prediction in Urban Transport," in *Proc. of Conference on Mobile Computing and Networking (MOBICOM'08)*, San Francisco, CA, September 2008, pp. 58–69.

[7] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, "Feedback Effects Between Similarity and Social Influence in Online Communities," in *Proceedings SIGKDD '08*. New York, NY, USA: ACM, 2008, pp. 160–168.