Distributed Content Curation on the Web

Zeinab Abbassi
Columbia University
zeinab@cs.columbia.edu

Nidhi Hegde
Technicolor, France
nidhi.hegde@technicolor.com

Laurent Massoulié
Microsoft Research - INRIA
Joint Centre, France
laurent.massoulie@inria.fr

ABSTRACT

In recent years there has been an explosive growth of digital content in the form of news feeds, videos, and original content, on online platforms such as blogs and social networks. We consider the problem of curating this vast catalogue of content such that aggregators or publishers can offer readers content that is of interest to them, with minimal spam. Under a game-theoretic model we obtain several results on the optimal content selection and on the efficiency of distributed curation.

1. INTRODUCTION

In recent years there has been an explosive growth of digital content in many forms, such as blogs, news feeds, original content and content propagated through online social networks. There is thus a need for recommendation and filtering of content, especially content that arrives in streams. Online social networks such as Google+, Twitter, and Facebook are rapidly turning into social reading or social content sharing networks, whereby each user shares content such as news, videos, etc., with her friends or followers. By resharing content, each user acts to filter items she receives or produces such that it is of interest to her followers. Indeed online aggregators serve to perform a similar function of content curation where content is chosen such that it matches with subscribers' interest.

Content curation can be quite complex when users (following or subscribing to the same aggregator) have differing interests and a limited budget of attention. By the latter we mean that users have limited time to sift through contents in their stream. The curation function by the aggregator then must take this into account by choosing a relatively small set of contents that followers would find most interesting. We address the scenario of content curation where a set of aggregators, or publishers, aims to optimize the set of content to publish such that the followers receive contents of interest to them.

We consider a very general setting with a set of aggregators or publishers, a set of followers or readers, and a set of contents or items of interest to the followers. Each reader has an intrinsic interest in a given item, represented through a quantified value. The problem then is: how should the publishers choose a limited set of items to publish such that readers receive contents of maximal value? As we will show, the problem of centralized optimization of this set of items is NP-complete, but admits approximation algorithms. We address the distributed content curation problem through a game-theoretic model where publishers are strategic agents with the aim of maximizing their utility, expressed in terms of feedback or incentives they receive from readers. In our analysis we aim to answer the following questions:

- Does distributed content curation based on selfish behavior of publishers converge to a pure equilibrium?
- How efficient is decentralized content curation? We define efficiency in terms of the price of anarchy, the ratio of social welfare under centralized curation to the sum of utilities achieved through distributed means.
- What is the rate of convergence to equilibria?
- How do the results change when readers are also strategic?

Our contributions. To the best of our knowledge our model is the first to address the problem of distributed content curation as a game. We show that with a certain form of feedback, or incentive, the distributed content curation game is efficient. In particular, we show that the game has a pure Nash equilibrium(NE) and that best-response dynamics converge to a pure NE. Further, we show that the price of anarchy is bounded by 2. We study the speed of convergence and show that in a polynomial number of best-response moves, players converge to a solution with an approximation factor of $2 + \epsilon$.

In the case of centralized content curation, we show that the problem is NP-complete. We further show that there exists a $(1 - 1/e)$-approximation algorithm for this problem.

Lastly we show that when readers are strategic in a certain sense, the price of anarchy is bounded by 2.

In the next section we present our model of content curation. We then analyze distributed content curation as a game in Section 3 and the centralized case in Section 4. We consider strategic readers in Section 5 and conclude in Section 7.

2. MODEL

We consider a scenario where each user connects to, or follows other users, and receives content such as videos, news updates, etc., posted by those aggregators or publishers. Our aim is to analyze content filtering behaviour whereby a set of users filter contents so as to optimize the content received by those connected to these users. As such, we consider a setting with a set of users that act as aggregators (or publishers), collecting and filtering contents from sources, and a set of users (or readers) that connect to these aggregators with the objective of receiving content of interest

to them.We will assume a set of exogenous content sources, that is, sources that have no strategic actions. Examples of such sources are news sources such as newspaper and magazines, owners of online videos, etc.

We further consider the more realistic setting where readers have a limited budget of attention, that is, they spend a limited amount of time in the act of content consumption. This translates to a limited number of publishers that the readers consult, and a limited number of items obtained when a publisher is contacted. Each user has some intrinsic interest in the items, represented by a value that a user attaches to each item. The problem then is how to allocate the limited number of links to friends and to select the limited number of items hosted at publishers so as to maximize the total value received by users. We now formally specify our model.

2.1 System model

Our system consists of a set $\mathcal{P}$ of P publishers or aggregators, a set $\mathcal{R}$ of R readers or users, and a set $\mathcal{C}$ of C contents or items. This can be represented as a graph $G = ((\mathcal{C}, \mathcal{P}, \mathcal{R}), E)$ of three types of nodes: content nodes, publisher nodes, and reader nodes, and of a set of directed edges. A directed edge from a publisher node to a content node indicates that this publisher selects this content to host. A directed edge from a reader node to a publisher node indicates that this reader follows this publisher; *following* in the social reading context has the meaning that a user receives contents posted by the publisher he follows. Define $\mathcal{F}_i$ to be the set of readers that follow publisher i, and $\mathcal{H}_j$ to be the set of publishers that reader j follows. The limited budget of attention then implies that a reader follows a limited number of publishers, or $|\mathcal{H}_j| \leq L$ for all readers $j \in \mathcal{R}$.

2.2 Publishing model

We assume that publishers post a set of items to their wall or stream and this set is accessible to their followers. Examples are news aggregator streams or the Twitter stream of a user. Let $\mathcal{C}_p(i)$ denote the set of items posted by publisher i and $\mathcal{C}_r(j)$ the set of items received by reader j, that is, $\mathcal{C}_r(j) = \{k : k \in \cup_{i \in \mathcal{H}_j} \mathcal{C}_p(i)\}$. The limited of budget of attention and filtering role of publishers implies that each publisher $i \in \mathcal{P}$ selects a set of items $\mathcal{C}_p(i) topost(orshare), with |\mathcal{C}_p(i)| \leq K$.

Readers have intrinsic interests in receiving certain items. Such interests are represented by a value; denote $v(j,k)$ to be the value of item k to reader j. We define the utility received by a user u_j as the total value of all items he receives: $u_j = \sum_{k \in \mathcal{C}_r(j)} v(j,k)$.

The publishers have no intrinsic interests in items, however receive a reward from their followers corresponding to the value of items posted. Specifically, publisher i receives a reward of $r_i(j,k)$ from reader j for posting item k. Note that these rewards can be in the form of feedback or incentives, through endogenous tools such as +1s, likes, retweets, or some exogenous form of monetary incentive. At this stage, we leave the type of this feedback quite general, only stipulating that its value be of the form shown in the analysis.

We now define the content curation problem with strategic publishers as choosing a set of items to be posted at each publisher such that the global utility is maximized. We study this problem in Sections 3 and 4, and in Section 5 we will consider the problem with selective readers.

Note that in the present formulation we consider a set of discrete items to be shared. In reality, once an item is *consumed*, it is no longer of interest and a user would seek other items. However, our formulation includes the general case where content is classified by topics and users have intrinsic value for certain topics. Each *item* k in our formulation then represents a stream of content of *topic* k.

3. CONTENT CURATION GAME

We first consider the case where the readers follow a fixed subset of publishers and read all items shared by the publishers they follow. The strategic publishers then select a set of items to post that maximizes their payoff. Each reader j is assumed to follow a fixed subset of publishers $\mathcal{H}_j$ and receives a utility u_j that corresponds to the items he receives from those publishers. The utility of each publisher i, r_i, is the sum of the rewards he receives from his followers for each item served. The form of these utilities and rewards will be made more precise below.

Each publisher chooses a set of K items to post that maximizes his total reward. Denote by A_i the set of feasible actions available to publisher i: $A_i = \{a_i \subseteq \mathcal{C} : |a_i| \leq K\}$. Each reader j receives a utility $v(j,k)$ for each item k that he receives. The reward, or price, corresponding to an item that he sends to publishers is divided among all publishers who post that item. This reward might be likes, re-shares, favorite markings, etc. in the various online social networks accordingly. The reward, or payoff, that a publisher receives for action a_i is thus as follows:

$$W_i(a_i) = \sum_{k \in a_i} \sum_{j \in \mathcal{F}_i} v(j,k)/n(j,k), \tag{1}$$

where $n(j,k) = |\mathcal{H}_j \cap \mathcal{P}_k|$ is the number of publishers through whom reader j receives item k, and $\mathcal{P}_k = \{i : k \in \mathcal{C}_p(i)\}$ is the set of publishers posting item k. This can be interpreted as the expected payoff received by the publisher. The tuple $\mathcal{G} = (\mathcal{P}, \{A_i\}, \{W_i(\prod_\ell A_\ell)\}$ now denotes our content curation game.

Note that W_i represent the private utilities of the agents. The social welfare is defined as the sum of utilities received by all readers:

$$\mathcal{W}(\mathcal{A}) = \sum_{j \in \mathcal{R}} u_j = \sum_{j \in \mathcal{R}} \sum_{k \in \bigcup\limits_{i \in \mathcal{H}_j} A_i} v(j,k),$$

where $\mathcal{A} = \{a_1, \ldots, a_P\}$ is the action profile of all users.

With the aim of characterizing the efficiency of distributed content curation, we consider the price of anarchy of the content curation game. The price of anarchy is defined as the ratio between the social welfare of an optimal allocation of items to publishers and that of the worst-case equilibrium. We will show that the price of anarchy is at most 2. Note that a Nash equilibrium of this game may be one of mixed strategy, where an agent selects an action according to some probability distribution. We will show that the content curation game admits at least one pure Nash equilibrium.

Let Ω denote an optimal action profile in the content curation game, that is, one that maximizes $\mathcal{W}$, the social welfare.

THEOREM 1. *Any Nash equilibrium of $\mathcal{G}$, the content curation game, results in social welfare at least half of the maximal social welfare:*

$$\mathcal{W}(\Omega) \leq 2\overline{\mathcal{W}}(A), A \in \mathcal{A},$$

where $\overline{\mathcal{W}}(A)$ is the expectation of $\mathcal{W}$ over the mixed strategy set A.

PROOF. Vetta [7] has shown that valid utility games have a price of anarchy of at most 2. It suffices to show that $\mathcal{G}$, the content curation game, is a valid utility game. A valid utility game has the three following properties:

(a) non-decreasing submodularity: the social welfare function must be submodular and non-decreasing,
(b) Vickrey condition: the utility of an agent is at least equal to the loss in the social welfare resulting from this agent declining to participate in the game,
(c) cake condition: the sum of agent utilities under any set of strategies should be less than or equal to the social welfare.

We now show that the content curation game satisfies these three properties.

(a) Since all item values $v(\cdot,\cdot)$ are non-negative, the social welfare function is non-decreasing. To show its submodularity, recall that $n(j,k)$ and $n'(j,k)$ are the number of publishers through whom reader j receives item k under strategy profile A and A' respectively. Note that since $a_i \subseteq a_i'$ for all i, $n(j,k) \leq n'(j,k)$. Let us now consider the increase in social welfare due to the utility of any reader j when item k is added to A and A'. If both $n(j,k)$ and $n'(j,k)$ are non-zero, user j's utility is not affected under either strategy and so the increase in social welfare is zero. If $n(j,k) = 0$ and $n'(j,k) > 0$, the social welfare has a non-zero increase when adding item k to A, but no increase when added to A' since reader k already receives the item under A'. Finally if $n(j,k) = n'(j,k) = 0$, the increase in adding item k to A and to A' is both $v(j,k)$. Summing over all readers, the increase in total social welfare due to adding any item k under A is not less than that under A'.
(b) When publisher i declines to participate in the game (denoted by action set $\emptyset_i$), the loss in social welfare as compared to when he selects action a_i is $\mathcal{W}(\{a_1,\ldots,a_P\}) - \mathcal{W}(\emptyset_i, a_{-i}) = \sum_{k\in a_i}\sum_{j:j\in\mathcal{F}_i, n(j,k)=1} v(j,k)$. The publisher's payoff, had he selected action a_i is $W_i(a_i) = \sum_{k\in a_i}\sum_{j\in\mathcal{F}_i} v(j,k)/n(j,k)$, which is greater than or equal to $\mathcal{W}(\{a_1,\ldots,a_P\}) - \mathcal{W}(\emptyset_i, a_{-i})$.
(c) It is easily shown that $\sum_{i\in\mathcal{P}} W_i(A_i) = \mathcal{W}(\mathcal{A})$.

□

3.1 Convergence to Equilibria

We now show that the content curation game converges to a pure Nash equilibrium. Rosenthal [6] has shown that any congestion game has at least one pure Nash equilibrium by providing an exact potential function for such games. Further, this implies that best-response dynamics converge to an equilibrium. We leave the proof details of the following theoreom where we provide an exact potention function, to the long version of this abstract.

THEOREM 2. *The content curation game is a congestion game. In fact, it is a potential game and any best-response sequence of actions will converge to a Nash equilibrium of this game.*

3.2 Convergence to Approximate Solutions

In the previous section, we showed that any sequence of best-response dynamics will converge to a pure Nash equilibrium. However the rate of convergence (in terms of the number of best-response moves) is not guaranteed to be polynomial. In this section, we study the rate of convergence of approximate (α-Nash) dynamics to approximately optimal solutions. In particular, we show that the number of approximate best responses by players before they converge to a solution within a factor $2+\epsilon$ of the optimal solution is bounded by a polynomial. In our analysis, we use the concepts of α-Nash Dynamics and β-nice games, defined in [1]. We defer the details of the description and proofs to the extended version of this abstract.

First, let us define approximate game dynamics and introduce some notation.

α-Nash Dynamics: An α-approximate best-response dynamics or α-Nash dynamics is a sequence of best responses by players in which each best response will increase the payoff of player (who makes the change) by a factor of at least α. In an α-Nash dynamics with liveness property, each player gets a chance to play a best response after at most T steps.

β-nice games: Consider an exact potential game Λ with potential function ϕ , and let Ω be the optimal solution. Let $\mathcal{A} = (a_1,\ldots,a_P)$ be a strategy profile of the players and let $\mathcal{A}_i'$ be a best response strategy for player i in strategy profile $\mathcal{A}$. The payoff of player i in strategy profile $\mathcal{A}$ is denoted by $\mathcal{W}_i(\mathcal{A})$ and each player wants to maximize its payoff. In this setting, in a strategy profile $\mathcal{A}$, for each player i with the best response strategy $a_i' \in A_i$ for player i, we let $\Delta_i(\mathcal{A}) = \mathcal{W}_i(\mathcal{A}_{-i}, a_i') - \mathcal{W}_i(\mathcal{A})$. We say that the game is a *β-nice game* if for any action profile $\mathcal{A}$,

$$\beta\cdot(\mathcal{W}(\mathcal{A}) + \sum_{i\in\mathcal{P}}\Delta_i(\mathcal{A},\Omega_i)) \geq \mathcal{W}(\Omega).$$

1-bounded jump condition: We say that a game satisfies *1-bounded jump condition* if for any action profile $\mathcal{A} = (a_1, a_2, \ldots, a_P)$, and any player i with best-response move a_i', and for every player i' the following two properties hold:

1. $\mathcal{W}_{i'}(\mathcal{A}_{-i}, a_i') - \mathcal{W}_{i'}(\mathcal{A}) \leq \mathcal{W}_i(\mathcal{A}_{-i}, a_i')$.
2. for every improvement action $a_{i'}'$ of player i', it holds $\mathcal{W}_{i'}(\mathcal{A}_{-i'}, a_{i'}') - \mathcal{W}_{i'}(\mathcal{A}_{-\{i,i'\}}, a_i', a_{i'}') \leq \mathcal{W}_i(\mathcal{A}_{-i}, a_i')$.

Awerbuch et al. [1] show that in order to prove the desirable result of this section, it is sufficient to prove that content curation games satisfy the above two properties. Next, we show that the content curation game satisfies the above two properties.

LEMMA 1. *The content curation game is a 2-nice game, i.e., if ϕ is the potential function, and Ω is the optimal solution, then for any action profile $\mathcal{A}$, we have $2\cdot(\mathcal{W}(\mathcal{A}) + \sum_{i\in\mathcal{P}}\Delta_i(\mathcal{A},\Omega_i)) \geq \mathcal{W}(\Omega)$.*

LEMMA 2. *Content curation games satisfy the 1-bounded-jump condition.*

As stated earlier, Theorem 5.4 of [1] and Lemmas 1, 2 imply the following corollary for the convergence of α-Nash dynamics to 2-approximate optimal solutions.

THEOREM 3. *Let $\frac{1}{8} > \delta \geq 4\alpha$. Consider a content curation sharing game Λ with and any initial strategy profile $\mathcal{A}_{init}$. Any α–Nash best-response dynamics with liveness property generates a profile $\mathcal{A}$ with total welfare $\frac{1}{(2+O(\delta))}OPT(\Lambda)$ in at most $O\left(\frac{n}{\alpha\delta}\log(\frac{\phi^*}{\phi(\mathcal{A}_{init})})\cdot T\right)$ steps.*

4. CENTRALIZED CURATION

We now study the centralized curation problem, where a central authority with complete information optimizes the set of items each publisher must post so that social welfare is maximized. For both results in this section, we defer proof details to the long form of the paper.

We first show that the problem is NP-complete by giving a reduction to the Set Cover problem.

THEOREM 4. *The maximum content curation problem is NP-complete.*

We then show that there exist $(1 - 1/e)$-approximation algorithms based on LP and greedy $(1/2)$-approximation algorithms by showing that the content curation problem is a special case of the separable assignment problem [2].

LEMMA 3. *The maximum content curation problem is a special case of the separable assignment problem, thus admits a $(1 - 1/e)$-approximation factor through an LP-based algorithm and a $(1/2)$-approximation through by a greedy algorithm.*

5. STRATEGIC READERS

We have thus far assumed that readers are not strategic, that is they follow a fixed set of publishers and read all the items published by them. However, due to their limited budget of attention, readers may only read a set L of items posted by publishers. The readers choose this set of L items strategically, such that their utility is maximized. Now, we consider the same utility model as in the previous sections, but with strategic readers that choose only L items to read. The optimization is still NP-complete[1].

From the game theoretic point of view, we can show that when followers are strategic as described above, the price of anarchy is 2. We can prove this by showing that the game is a valid utility game[1].

6. RELATED WORK

To the best of our knowledge, our model is the first to address the problem of content curation by a set of aggregators for the optimization of utility of a set of readers. However, similar models of optimization and game theory have been considered in different contexts.

The content curation game is related to a previously studied game called the market sharing game [4]. In a market sharing game, there are a set of players (agents), each playing a subset of markets, and we are given a bipartite graph between the markets and the players indicating which markets are eligible to be played by each agent. A generalization of market sharing games is the distributed caching game where the strategy space of players is more general than playing a subset of markets [2]. Content curation games are different from both types of games in that in content curation games, there are three parties in the game: the players (or publishers), the items they post (corresponding to markets), and a third party that is the set readers who follow a subset of publishers. Neither of these three games is a special case of the other, and in particular content curation games are a generalization of "uniform market sharing games" [4].

Our work is also related to another line of work that explores incentives in user-generated content systems where users are strategic. In particular, in [5], the authors study the problem of strategic news posting in online social networks. They model users as either greedy or courteous in their posting strategy. They analyze these two models on random graphs from a game theoretic point of view. They find that high quality information spreads in the network if users are greedy. Through simulations on Twitter data they show the same observation when users are modeled as courteous. In our work, rather than information spread, we are interested in maximizing the utility of readers where they have differing interests in content.

Another line of work in the area of user-generated content models ranking mechanisms in a game theoretic setting, where utility is defined in terms of the attention (exposure) the contributor or their content receives. In this setting, generating higher quality content is assumed to be costlier. The objective of the system is to elicit high quality and high participation in equilibrium [3]. Our model is more general where we consider content aggregation, not modification of content quality.

7. CONCLUSION

We have considered the problem of content curation by a set of publishers aiming to maximize global utility of a set of readers. We show that the centralized optimisation, while being NP-complete, can be reduced to a separable assignment problem, thus admitting a $(1 - 1/e)$ approximation algorithm. We model distributed content curation as a reader-publiser game and show that the price of anarchy is 2. When in addition the readers are selective in the items they choose, we show results on the price of anarchy. The complete version of this extended abstract will include proofs of all our results and more detail on the case with strategic readers.

8. REFERENCES

[1] B. Awerbuch, Y. Azar, A. Epstein, V. S. Mirrokni, and A. Skopalik. Fast convergence to nearly optimal solutions in potential games. In *ACM Conference on Electronic Commerce*, pages 264–273, 2008.

[2] L. Fleischer, M. X. Goemans, V. S. Mirrokni, and M. Sviridenko. Tight approximation algorithms for maximum separable assignment problems. *Mathematics of Operations Research*, 36(3):416–431, 2011.

[3] A. Ghosh and P. McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 137–146, New York, NY, USA, 2011. ACM.

[4] M. X. Goemans, E. L. Li, V. S. Mirrokni, and M. Thottan. Market sharing games applied to content distribution in ad-hoc networks. In *MobiHoc*, pages 55–66, 2004.

[5] M. Gupte, M. Hajiaghayi, L. Han, L. Iftode, P. Shankar, and R. Ursu. News posting by strategic users in a social network. *Internet and Network Economics*, pages 632–639, 2009.

[6] R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

[7] A. Vetta. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *FOCS '02*, page 416. IEEE Computer Society, 2002.

[1]We omit the proofs of this part for lack of space, but will include them in the complete version of the paper.