Tutorial: Incentivizing and Coordinating Exploration

Robert D. Kleinberg∗  Aleksandrs Slivkins†

February 2017

An increasing variety of platforms and markets rely on the activities of self-interested agents to explore a space of alternatives by engaging in the (often costly) process of acquiring information about those alternatives. The cost of exploration may be direct, such as paying to interview job candidates prior to making a hire or to visit colleges prior to deciding on a school, or it may be an opportunity cost, such as the myopically suboptimal decision to try a new restaurant to discover whether it is more appealing than one’s current favorite. Either way, when individual agents bear the full cost of their exploration but are not guaranteed to derive the full benefit, there is a potential for market inefficiency and, accordingly, a need for research on mechanisms that mitigate the inefficiency. The past four years have witnessed a surge of interest in research on such mechanisms, as people from various fields have attacked these issues from many different angles. Some common themes are the use of models based on multi-armed bandits and a focus on information asymmetry between the platform and the agents. Prototypical applications range from platforms for ratings and recommendations, to startup acquisitions, to medical trials.

Our scope. We survey recent work in computer science, economics and operations research on incentivizing and coordinating exploration. All models we survey share the following properties. A principal (i.e., an algorithm) interacts with self-interested agents whose actions may reveal information not previously known to themselves, the principal, or other agents. The choice of information-revealing actions is directly controlled by the agents. The principal can only influence the agents via signals (e.g., action recommendations) and/or monetary transfers. Principal and/or agents can learn — aggregate and subsequently use the new information revealed by agents’ actions. Following much literature in economics and computer science, the principal has the power to commit: she commits to using a particular algorithm for interacting with the agents, announces this algorithm to the agents, and the agents believe she actually uses this algorithm.

Absent incentives, these models reduce to various multi-armed bandit problems (Gittins et al., 2011; Bubeck and Cesa-Bianchi, 2012). In these problems, an algorithm repeatedly chooses from a fixed set of alternatives (a.k.a. arms), collects rewards for the chosen arms, and receives little or no feedback about the other arms it could have chosen instead. This is a clean, abstract model for an ubiquitous tradeoff between exploration and exploitation: making optimal decisions using information collected in exploration. Rather than a two-way tradeoff between exploration and exploitation, we consider a three-way tradeoff between exploration, exploitation, and agents’ incentives. Indeed, agents are typically modeled as short-lived and/or myopic, and therefore have a strong preference for exploitation vis-a-vis exploration, whereas the principal typically strives to balance the two.

The main distinction within our scope is, who learns: the principal or the agents? When agents learn, but the principal does not, this constitutes a mechanism-design counterpart to classical economic models of search and matching, studied in Kleinberg et al. (2016). Here consumers engage in exploration to zoom
in on better alternatives, and a principal can coordinate this process to make it more efficient. Scenarios when the principal learns, but agents’ learning is de-emphasized, arise in recommendation systems, and are studied in (Kremer et al., 2014; Che and Hörner, 2015; Frazier et al., 2014; Mansour et al., 2015, 2016; Bahar et al., 2016; Bimpikis et al., 2017).

Most of the above papers, with notable exceptions of (Frazier et al., 2014; Kleinberg et al., 2016), do not allow monetary transfers. Several papers consider rewards without time-discounting (Kremer et al., 2014; Mansour et al., 2015, 2016; Bahar et al., 2016) and are intellectually connected to the literature on regret-minimizing bandit algorithms. Some others study time-discounted rewards (Frazier et al., 2014; Kleinberg et al., 2016; Bimpikis et al., 2017), and are close in spirit to the work on Bayesian formulations of multi-armed bandits, and particularly to Gittins algorithm (Gittins and Jones, 1974).

The literature within our scope makes several other notable modeling choices (and studies both sides of each choice): whether agents can observe other agents’ actions, whether agents’ rewards depend on other agents’ actions, and whether reward distributions allow for Chernoff-like concentration bounds.

**Motivating applications.** Our setting is broadly applicable to the numerous platforms that collect ratings and/or make recommendations about a space of alternatives: movies (e.g., Netflix), restaurants (e.g., Yelp), vacations (e.g., TripAdvisor), products (e.g., Amazon), driving routes (e.g., Waze), doctors (e.g., SuggestADoctor.com), and so forth. In fact, the ability to make high-quality recommendations is an essential part of the value proposition for the corresponding businesses. Second, our setting is relevant to auctions and matching markets whose participants face substantial uncertainty about their options and/or their own values, and incur costs for acquiring such information. For example, efficient information discovery (or lack thereof) is an important issue in various real-life matching markets for jobs and other positions in the U.S., such as college admissions, medical residency admissions, and job markets within specific academic disciplines. Third, well-coordinated exploration is crucial in large-scale acquisitions under uncertainty, such as start-up acquisitions and real-estate purchases. Finally, incentivizing large-scale participation (while mitigating selection biases) is a major issue in medical trials, especially for wide-spread diseases and inexpensive treatments.

**Closely related work not in our scope.** Several models in prior work combine exploration and incentives, and lie just outside of our scope. In fact, they can be seen as “one-step deviations”, when we keep all the major tenets in our scope except one:

- Similar models but without a principal are known as strategic experimentation (Bolton and Harris, 1999; Keller et al., 2005).
- Mechanisms for exploration where information acquisition is not controlled by agents have been studied in various settings: dynamic auctions (e.g., Athey and Segal, 2013; Bergemann and Välimäki, 2010; Kakade et al., 2013), ad auctions (Babaioff et al., 2014; Devanur and Kakade, 2009; Babaioff et al., 2015), and human computation (Ghosh and Hummel, 2013).
- Similar models where information is aggregated rather than acquired — namely, when the principal aggregates information that is already known to agents — have been studied in dynamic pricing (e.g., Kleinberg and Leighton, 2003; Besbes and Zeevi, 2009; Badanidiyuru et al., 2013).

The design of “information structures” — essentially, policies for revealing information to agents — has been an important line of work in theoretical economics starting from Kamenica and Gentzkow (2011) and Bergemann and Morris (2013); see Dughmi and Xu (2016) for an algorithmic angle. Our focus is on designing information structures with a particular emphasis on exploration.

**Structure of the tutorial.** The proposed tutorial consists of two segments. One segment covers the work involving time-discounted rewards and monetary transfers, focusing on the material in (Frazier et al., 2014; Kleinberg et al., 2016) and drawing strong intellectual connection to Gittins algorithm. The other segment
considers scenarios with no time-discounting and no monetary transfers, focusing on the progression of 
papers (Kremer et al., 2014; Mansour et al., 2015, 2016; Bahar et al., 2016), and particularly on the results 
in Mansour et al. (2015). While these papers model beliefs and incentive-compatibility using Bayesian 
priors, their approach to algorithm design is essentially non-Bayesian, following the rich literature on regret-
minimization.

Tutor biographies

Bobby Kleinberg is an Associate Professor of Computer Science at Cornell University. He was also a 
researcher at Microsoft Research New England from 2014 to 2016. His research in general pertains to the 
design and analysis of algorithms, and their applications to economics, machine learning, networking, and 
other areas. Prior to receiving his doctorate from MIT in 2005, Kleinberg spent three years at Akamai 
Technologies, where he assisted in designing the world’s largest Internet Content Delivery Network. He is 
the recipient of a Microsoft Research New Faculty Fellowship, an Alfred P. Sloan Foundation Fellowship, 
and an NSF CAREER Award. His research has received the best paper awards at ACM EC 2010 and 2014.

Alex Slivkins is a Senior Researcher at Microsoft Research New York. Previously he was a researcher at 
MSR Silicon Valley in 2007-2013, after receiving his Ph.D. from Cornell in 2006 and a brief postdoc at 
Brown. His research interests are in algorithms and theoretical computer science, spanning machine learn-
ting theory, algorithmic economics, and networks. Alex is particularly interested in exploration-exploitation 
tradeoff and online machine learning, and their manifestations in mechanism design and human computa-
tion. His work has been recognized with the best paper award at ACM EC 2010, the best paper nomination 

References

preliminary version has been available as a working paper since 2007.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mecha-

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computa-


Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic recommendation systems. In *16th ACM Conf. on 


liminary versions have been available since 2006, as *Cowles Foundation Discussion Papers #1584* (2006), #1616  
(2007) and #1672(2008).

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal 


