Application to Digital Economics and AI Tutorial

Jasmine Yang
Columbia Business School
Email: yyang24@gsb.columbia.edu

I am currently a sixth year PhD candidate at Columbia Business School specializing in quantatative marketing. My dissertation, entitled "creator content production and promotion decisions on online video platforms", is reviewed by a committee comprising of Oded Netzer (co-advisor), Andrey Simonov (co-advisor), Olivier Toubia, and Hortense Fong. My dissertation contains a set of papers all related to understanding the growing role and influence of the creator economy and digital market.

I am intrigued by the interplay among creators, viewers, advertisers, platforms, and regulators in digital markets, as they collectively represent a complex network of interactions (e.g., sometimes with misaligned incentives). Across several papers, I have studied the decisions of both the demand-side (consumers) and supply-side (creators and platforms) of the digital market. To address these topics, my research applies diverse methodologies such as computer vision, causal inference, high-dimensional/unstructured data, structural models, experiments, applied game theory, and Bayesian statistics.

More specifically, I study the production and promotion decisions of content creators and the impact of platforms' content policies by leveraging new types of data, including high-dimensional, unstructured data (e.g., video content that can be analyzed through images, texts and audio) and novel, high-frequency data (e.g., self-collected streaming and viewing decision data by monitoring livestreams of 30,000 streamers and their viewers every 15 minutes for nearly a year). In terms of research context, I find Twitch and YouTube useful research settings to study the digital economy, as their market structure (e.g. competition among creators and frequent viewer substitution across streams and/or channels) as well as the diversity of products for consumption (e.g., streams of video games, user-generated video content), constitute ideal quantitative marketing laboratories to study creators and platform decisions and the impact on the overall content market.

My research in the area of the creator economy and digital market can be categorized into two main subcategories.

Leveraging unstructured data and modeling to understand demand and supply in digital markets

The digital economy often involves unstructured data in the form of videos, audio, image, and text. Hence, to understand the drivers of this economy and incorporate the features of this economy into traditional models, one needs to extract features from unstructured data and often reduce the dimensionality of such data. In the first two chapters of my dissertation, I unite two papers

under the joint theme of providing a structured framework of thinking how content creators (e.g., online influencers, designers, advertisers) should select thumbnails (reduced sized preview images or clips) for their video content. While this question has significant implications for the creator economy, suprisingly, it has received scant attention in the literature. To address this question, I involve two different data sources and multiple different approaches involving both video analysis, computer vision, advanced large language models (LLMs), creation of video platform for realistic experimental design and structural model.

In the first paper, I propose a video mining procedure and construct unique, theory-based metrics based on interpretable features extracted using computer vision, text mining, deep learning, and advanced large language models (LLMs) to study how the thumbnail of a video (relative to video content) affect viewers' subsequent video reactions (e.g., views, watchtime, engagement and preference for the video). Drawing on behavioral theories such as expectation-disconfirmation theory and Loewenstein's theory of curiosity, I postulate that thumbnails can either serve as expectation-based reference points (which shape viewers' expectations for content before they click to see a video) and/or as informational reference points (which build anticipation for specific moments within the video). Using secondary analysis of YouTube data and experiments, I find a trade-off between "clickbait thumbnails", which maximize viewership at the cost of engagement, and content representative thumbnails, which maximize engagement. Additionally, choosing thumbnails with some content difference with the video can increase watchtime through information gap and curiosity. I rationalize the roles through which thumbnails may affect video reactions and propose optimal thumbnail designs to enhance creators' content promotional strategies.

In the second paper, I address the limitations of secondary analysis in my first paper by creating my own video streaming platform (CTube, a simplified version of YouTube), which allows me to (1) collect high-frequency clickstream data to track individual-level video choices and watchtime decisions, and (2) exogenously randomize the thumbnails that viewers see to start a video to assess the causal effect of thumbnails. Using click-stream data, I build and estimate a Bayesian learning model, in which thumbnails serve as priors for the video content (characterized as multi-dimensional and correlated video topic proportions) and frames in the video serve as signals to update viewers' beliefs as they continue to watch the video. I use the model estimates to explain the reference processes through which thumbnails affect video watchtime and evaluate counterfactual video frames to recommend thumbnails that enhance video performance for creators and platforms.

While the two papers are very much related substantively, the first paper primarily leverages computer vision, machine learning, and LLMs, the second model leverages sophisticated experimental design to create a video website that the researcher can fully control, a structural Bayesian learning model, and a challenging Bayesian estimation approach with a large parameter space.

Understanding market structure, production and consumption decisions using big data and modeling

In the last two chapters of my dissertation, I study the production decisions of content creators, viewers' preference for creators and their content, and the implications of platform policies on content production and consumption. I find Twitch as an ideal setting to study these questions - in this market, streamers face high competitive pressure from other streams (or other entertainment options of the viewers), and their incentives may not be aligned with that of the platform which also relies on streamers to supply content to viewers.

In order to gain insights into the business context of the creator economy, during one of the summers of my PhD studies, I interned as a PhD Applied Scientist in Twitch's central science team. I developed several research ideas grounded in Twitch as the empirical context. In one dissertation essay, I collected a novel high-frequency dataset on streamers' broadcasting decisions and viewers' watching decisions by monitoring livestreams of 30,000 pre-selected streamers (based on a 1-month pilot study) and their viewers every 15 minutes for nearly a year. Running network analysis separately for streamers and viewers, I find evidence of production inefficiencies in this market (e.g., extraneous content on the platform that few viewers watch, and misallocation of production decisions of streamers even for content that is watched). I argue that these inefficiencies stem from two types of streamer incentives: "streamers-as-firms" who focus on profits may overinvest in production of popular content with high expected returns, and "streamers-as-consumers" who produce content with high personal preference may not produce what viewers want to watch. I use detailed individual-level viewership data to estimate viewers' tastes for streamers vs. content. Using high-frequency data on streamers' production decisions, coupled with demand estimates, I then estimate weights on streamers' two incentives to quantify market inefficiencies and evaluate counterfactual policy prescriptions for welfare improvement.

In another, I use part of the same high-frequency dataset with unstructured data and econometrics tools to study a specific policy shock (e.g., Twitch's ban of production of gambling sites lacking U.S. license) which shifts the production decisions of streamers. Using network analysis to construct valid treatment and control groups, I estimate the effect of the policy on viewership through which I understand whether platforms (e.g., care about profits) may have misaligned incentives with policymakers (e.g., concerned about consumer protection) and derive implications of the policy ban.

More recently and why this tutorial is important: Most recently, I am starting to examine the potential of large language models, such as ChatGPT, to foster empathy with consumers. Specifically, I am interested in whether these models can be employed to semi-automate the generation of insights by identifying consumer needs, motivations, and emotions. This tutorial aligns with my broader interest in digital markets and it can offer particularly valuable insights for this recent project of mine in terms of exploring ways in which artificial intelligence can enhance and expedite decision-making processes while simultaneously considering potential ethical concerns and costs brought by the rapid advancement of these technologies. This includes discerning when to deploy these technologies and when to engage human intervention to ensure, for example, the insights generated do not fall into the uncanny valley. Engaging with fellow researchers in this field through this tutorial will be greatly beneficial for my research in general and my understanding of the advantages and disruptions that artificial intelligence may potentially introduce to digital markets.