Data science for leveraging user perception

Data driven pipelines in the matters of human perception

Sagar Joglekar

Department of Informatics, Faculty of Natural and Mathematical Sciences

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

March 2019

To Geetika, Medha and Chanda ... The three formidable pillars of my life

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. The dissertation covers contributions for journals and conferences where I was the main contributor and primary investigator.

Sagar Joglekar
March 2019

ACKNOWLEDGEMENTS

And I would like to acknowledge ...

ABSTRACT

The idea of wisdom of the crowds has been tested time and again when it comes to applications like recommendations of items or places, which inherently taps into the likes and dislikes of the crowds. Interesting convergence has been shown to emerge when insights are driven by usage patterns of large number of people.

With the ever pervasive nature of the internet, we as a society have started treating the online spaces not only as a tool to access information seamlessly with atomic transactions, but also as a natural extension of the self. We spend more time than before as a part of a larger networked community, exchanging thoughts, debating ideas, expressing creativity, empathy and sometimes seeking help. The presence of these extended set of interactions over the web, implies that an intangible yet essential affective component is now driving our behaviour when we empathise with people online, or interact with creative or aesthetically pleasing content. At such a juncture, I ask the question "Can we build pipelines that look at these interaction patterns, to understand and utilize the perception driven behaviours of humans on the web?"

In this dissertation I explore the idea of "Perceptions of the crowds" through two case studies. I build systematic pipelines that quantify human perceptions in both cases interpreting interdisciplinary ontologies. In the first study I analyse on-line formal networks where interactions between humans are purely with the aim of helping each other. I develop frameworks to abstract out the structure of these conversations in computable structures. I then delve into investigating presence of supportive phenomenon by finding discriminative local structures in these abstractions. I reason about these local structures using established cross disciplinary theories. This informs my analysis about the nature of peer to peer support in these communities and paves the way to do actionable intervention in the area of peer support in online networks.

In the second study, I investigate utility of perceptions of physical spaces, by developing a pipeline that capitalizes on perception of urban aesthetics at the crowd scale, to develop intervenable insights around urban beautification. I propose a deep-learning driven framework, which is able to quantify the perception of intangible abstract qualities like ‘beauty’ through

a crowd sourced rating of google street view images. I show that a general pattern of beauty in urban spaces can be learnt through a crowd sourced opinion and deep learning models. I further develop a generative model to simulate beautification of urban spaces. Through a detailed literature review of the field of urban design, I develop a measurement framework which can provide insights into the predictors of urban beauty on a case by case basis using well known urban design metrics. I validate the value of these metrics through expert survey and validate the interventions through crowd sourced perception experiment.

As an over arching goal, through this dissertation, I try to build a repeatable framework for data driven pipelines that can capitalize on data that contains human perception signals. I propose that building custom representations that bind computational metrics with inter disciplinary ontologies is a necessity to build useful pipelines.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

One should never try to prove anything that is not almost obvious - Alexander Grothendieck

1.1 Introduction

We live in a world where information is being bombarded on our cognitive faculties from all sides, at all times. The internet is a continuous stream of information and each source is fighting with the other to get a piece of our attention budget. With the advent of machine learning and big-data sources, building systems that predict actions as a response to perceptual triggers is the bread and butter of many companies. The use cases may range from understanding which advertise made a visitor do an unscheduled purchases on amazon, or which string of music tracks recommendations maximized a users time on a particular music platform. But in the end it all boils down to understanding what triggers result in human action or lack there of[71]. Nonetheless the systems that surrounds a human interacting with the internet, are all figuring out the best triggers which are perceived by the human as worthy of attention. The term "Attention Economy"[11] was actually coined for this very reason. In the words of Matthew Crawford "*Attention is a resource, a person has only so much of it.*"[9]. We live in the age of distraction, and more often than not, our perceptions are guiding our actions, than our cognitive processes. Several studies have shown that engagement is almost always a game of stimulating our most basic urges, such as dopamine hits, presence of faces or simply arousal of emotions to increase the working memory.[3, 29][66][70]. An interesting side effect of dwindling attention budgets is the emergence of more formal topical spaces on the internet. The ever pervasive nature of the internet allow these formal spaces to function almost like physical communities, with moderated and effective peer to peer exchange of thoughts, ideas and empathy[26, 36, 73].

In such an environment, as computer scientists, it is worth asking the question: **Can we develop pipelines that can look at the structures in the data to quantify how humans perceive their sphere of interaction? Can we use that perception to design impactful interventions towards well being of the users, on and off the internet?**. This question can be decoupled and decomposed into two fundamental problems:

1. How do human perceptions manifest in data?
2. How do we quantify these perceptions?

These two questions are going to be the guiding principles of my dissertation.

But first of all, we need to clarify the relation between perception, affects and intervention and to do so we should try and understand each of these terms separately as a part of the native field's context.

1.2 Perception and Affect

This paragraph needs to be refined with more literature that convinces the reader that perception -> emotion. And any knowledge extraction pipeline inherently links the two together. In this dissertation, I would try to build frameworks to capture human perceptions in the realm of human to human interactions and subjective aesthetics. The utility of such an attempt, can only be justified if there is a real link between how humans function at the most fundamental cognitive level and how they perceive the intangible, including the aesthetic. There has been an ongoing effort to unravel this link, through psychological, neuro-evolutional and philosophical arguments. I will try to gain inspiration from them, but a detailed critique is beyond the scope of my dissertation and expertise. ***Affect (APA definition):*** *any experience of feeling or emotion, ranging from suffering to elation, from the simplest to the most complex sensations of feeling, and from the most normal to the most pathological emotional reactions.*
Perception (APA definition): *the process or result of becoming aware of objects, relationships, and events by means of the senses, which includes such activities as recognizing, observing, and discriminating. These activities enable organisms to organize and interpret the stimuli received into meaningful knowledge and to act in a coordinated manner.*

Emotions or 'affects' and perceptions have long been discussed in the psychology, neuroscience and philosophical literature. Emanuel Kant in his prolific work, first discussed the utility and the philosophical reasoning behind presence of affects or emotions[31]. In his opinion, emotions are pre-cognitive involuntary states, termed as "mere perceptions of

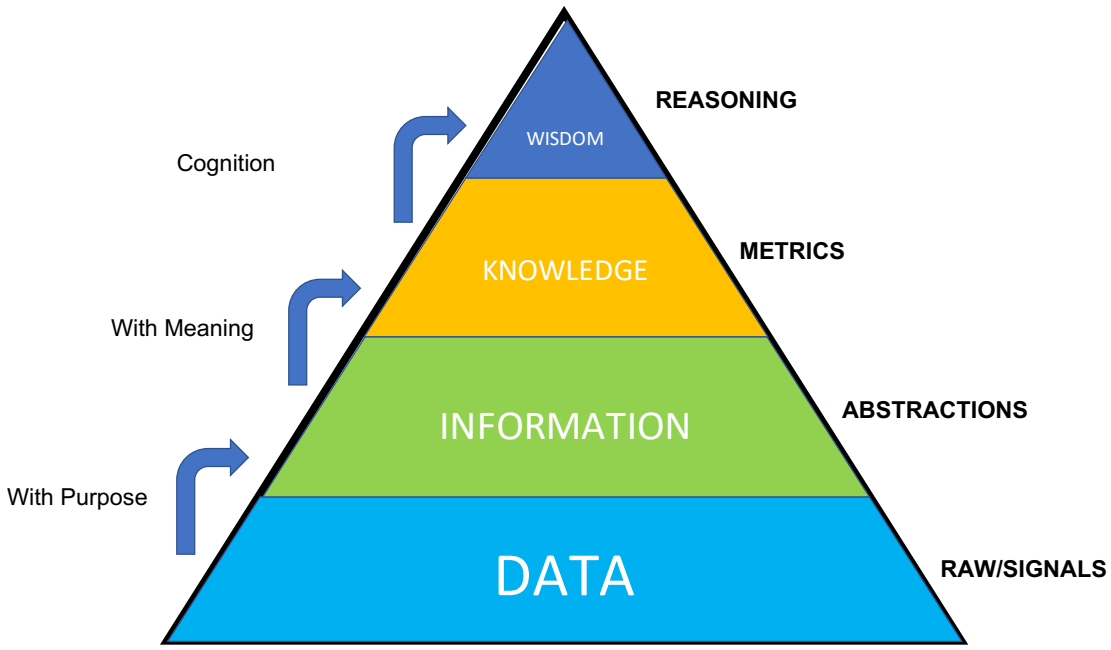


Fig. 1.1 The DIKW pyramid

unspecified bodily states"[5]. But that does not mean they don't influence our deepest level of well-being. The link between affect and perception has also been explored in several cases. An argument to link perception of affect arousing aesthetics was made by Perlovsky[55], where they propose that the phenomenon of affects arousing from aesthetics, comes from a fundamental human need enrich the knowledge bases about real world. An unexpected thing of structure, satisfies this need at some level and we perceive it as aesthetically pleasing. Another recent study by Zadra et.al[82] evaluated the relation between visual perception and emotions. They demonstrate that the conventional assumption of the disentangled functioning of perception and affects is not necessarily true. Humans are quite susceptible to perceiving different realities based on different aroused affects.

So the community is still unclear, but there seems to be a theory that

1.3 Intervention and the DIKW model

The reflection to find the answers the questions, ties us back to one of the fundamental frameworks about operations on data, that is the **Data , Information, Knowledge and Wisdom** model[61].In this dissertation I posit and demonstrate through case studies, that for reasoning about perceptions, you need to interpret the metrics that contain contextual knowledge about the abstractions you are working with through the lens of the target ontology. Debating the nature of this ontology is not the purpose of this dissertation, but I show that 'a'

solution can be reached, if systems are designed to interpret metrics by anchoring around an ontological bias.

In this model, the most base layer consists of the pure form raw **data or signals** that come from a source. If we are measuring perceptions of human, this source needs to be tied back to humans in some way. The data needs to be generated as a result of some human-human interaction or as a result of deliberate human input as a form of response to their perception of some event.

The **information** layer is the result of the fact that any process done on the data is with a sense of purpose or an end goal. For example, if the goal is to understand how humans exchange messages at times of distress, you would most certainly need to express the raw information about sender and recipient of messages into some form of a networked abstraction. The abstraction preserves the organization of data, but at the same time allows information to be operated on. As a result, almost always the output of this process is some form of a data abstraction. In the next stage of processing, you need to attach some meaning to the patterns in the information to extract

knowledge. This For example, if you need to know the most popular user who has exchanged the most messages with others among a network of users exchanging messages; you would look for the most central user in the network. In this particular case, the metric of centrality dawns the meaning of popularism in the context of our message network.

The final layer needs a cognitive process and an ontological framework, to extract actionable insights, which we can call **wisdom**, from the knowledge. By classical definition of ontology, it defines properties of and relations between objects. For this very reason, these ontological frameworks need to be originating from the field of intervention. In our example, a human analysing data needs to get some insights about the dynamics of popular users which requires the human to consume the metrics and derive insights. However the insights need to be grounded in psychological ontological framework, for us to derive certain trends in behavioural dynamics of this network. Figure 1.1 shows an illustration of the adopted version of Rowley's DIKM model.

1.3.1 Data

As discussed, this dissertation is about developing frameworks around quantification of human perceptions, such that we can do impactful interventions from this approach. And the most base level of this pyramid is the data that the frameworks would work with in order to progress on these lines. I work with two forms of data, textual data and image data, to understand two separate forms of perceptions.

Textual Data

The first case study of this dissertation focusses on is online support communities. It has been shown through several studies in medical informatics, that these communities play a very important role in providing support and respite in times of distress. The communities are especially helpful when it comes to people suffering from long term illnesses or mental health issues. To understand how users on these communities perceive social support, I work with data acquired from online health forums, where users share, give support and ask for support. I look at communities that deal with long term conditions like Lung illnesses, and communities where mental health patients seek support.

Image data

The other facet of my work looks for quantification of how we perceive physical spaces. I work with google street view data, and the aim of this work is to understand, through crowd sourced methods, how people perceive the sense of beauty in urban areas.

1.3.2 Abstractions

As mentioned before, the act of aggregating information from data, almost always involves building organized abstractions. In case of the first study with textual data, I incorporate user information into the textual data to build organized networked structures, which can then be evaluated using graph theoretic methods. To understand textual patters, I use the language embedding models which would be discussed in the chapters to come. While processing image data, I use several segmentation techniques to group semantically similar pixels together. I also use several state of the art object and scene detection to extract semantic information from an image. A more detailed discussion of these abstractions would be done in the later chapters.

1.3.3 Knowledge

For extracting knowledge, we need to first associate meanings to certain computable metrics that we obtain from the abstractions. As discussed in the previous example, it could be as simple as associating the property of "popularity" to the metric of centrality. In my case, I develop several of these metrics for the two studies, some based on intuitions which i test validity for, and some based on extensive literature survey. To give an example, I develop the concept of anchored triads, which combines local interaction motifs with the role of a user in a supportive conversation to understand how these conversations evolve.

1.3.4 Wisdom

Finally the wisdom in my case, is simply the philosophical, quantitative and qualitative discussion about what perception did we actually capture from these progressive operations. To answer the questions posed by this layer, we need to hold on to **an ontological framework** which grounds the metrics in the field of human interaction and planned intervention. This ontological framework that deems meaning to the structure in data, needs to be achieved through cross disciplinary literature review. In case of social support, what metrics combined with the understanding of human behavioural ontology, give rise to signatures of social support? Could those be found in sociological literature that distils social ties as interaction motifs? Or can that be found in pure statistical understanding of networks and formation of edges? And in the end, can we provide valuable interventions in the scenarios where humans are asking for help online. My dissertation navigates similar questions to arrive at 'an approach' to quantify and intervene using signatures of human perceptions from data.

1.4 Research hypothesis

The global hypothesis of my dissertation as described before comprises of asking the two aforementioned questions: **How do human perceptions manifest in data?** and **How do we quantify these perceptions?**. But these questions are quite open ended, and answering them in a generalized manner seems impractical. For this reason, I need to first contextualize my work in the realm of practical applications. To rationalize the pursuit of these questions, I choose application driven case studies, which allow me the luxury to do a data driven exploration with a final goal of designing intervenable frameworks. Across my investigations, I follow the DIKW framework layer by layer, by distilling actionable wisdom from data. Through the two case studies , my work touches a diverse set of data science tools which range from complex networks formulations to generative adversarial models of human perception.

My research explores the following global hypotheses using the two case studies.

H1: Data about human interactions from the web can be used to capture signatures of perceptual processes

H2: Human interactions data from the web can be used to quantify subjective attributes of the real world

H3: These quantifications can be effectively used to drive intervenable insights to improve perceptions of the said humans.

Human interactions data from the web can be used to capture peculiar formats of human conversations

Ontology in a philosophical context is a branch of metaphysics, which asks the questions like "What exists?", "What comprises an object? ", "What hierarchies of object classes are present?", "What different entities form a higher entity?". Naturally this framework was suitable in information sciences to describe datum as hierarchies or taxonomies of other objects.

1.5 Thesis overview

In the first study, I examine the structure behind how people seek engagement on-line. I discover that humans are very limited by their attention budgets, and engage with informal social networks in a very atomic and primacy driven way, which makes prediction of engagement using certain attributes of the content very feasible. But on the contrary, in more formal social networks, where the aim is to exchange knowledge and support, the network evolves around certain key elements of perception of support and helpfulness, which are well explored in the real world communities. On the journey to unravel these traits, I develop techniques to abstract out online formal conversations and develop a models for detecting supportive conversations on the web. I discuss the utility of such a model and draw parallels with the offline model of community support.

In the second study, I investigate utility of perceptions of real world places through a crowd sourced rating of google street view images. I develop models to extract the perception of the crowds using data driven inference methods. I show that a general pattern of beauty in urban spaces can be learnt through a crowd sourced opinion and based on this finding, I develop a generative model to simulate beautification of urban spaces by using deep learning. I validate the quantification of perception of real-world beauty using crowd validation. I contribute a way to use computer vision techniques to abstract out beautification process into explainable metrics used by architects and urban planners. The final contribution is a demo

web application, that allows practitioners to examine and validate the utility of such a end to end system that captures citizen perceptions for urban design.

1.6 Original contributions

1.7 Outlook

1.7.1 Notes

Some points to cover in introductions

- Kant's theory of emotions , and how stimulus -> cognition -> emotion and action links everything together
- importance of capturing affects (desire to influence behaviour)
- There have been attempts to capture this , and how the plutchik's frameworks has helped in distilling certain emotions and capturing them from different forms of media
- this dissertation takes an different approach, in which we try to find signatures of affective responses seen online
- the dissertaion works in two parts, the first part looks at how support can be detected from structures in coversation
- part 2 looks at how the affect of beauty impacts how citizens perceive cities and can we capture that affect to improve urban design process

CHAPTER 2

ATTENTION BUDGETS

2.1 Introduction

In the last few years, we have seen the introduction of a new form of user-generated video, where severe restrictions are placed on the duration of the content. High profile examples include Vine, which allowed users to create videos up to 6.5 seconds long; Instagram, which introduced videos up to 15 seconds duration; and Snapchat, whose videos are officially limited to 10 seconds and are deleted after 24 hours. Although most user-generated video platforms have placed some form of limit on the duration or size of videos (e.g., YouTube had a 10 minute limit, which has since been softened to a 'default' limit of 15 mins[1]), the extremely short duration time limits of Vine etc has led to the coining of a new term: *micro videos*. Some media commentators have argued that the restrictions imposed by the micro video format could fundamentally change the way we communicate [22]. Indeed, it has been argued that Vine has had a significant cultural impact far beyond its user base, generating several widely shared memes in its short lifetime[2].

At the same time, as the format is still very new, virtually all major micro video platforms are experimenting with the format, making significant changes in the last year. For instance, Instagram extended the limit from 15 seconds to 1 minute[3]. Vine is undergoing a major overhaul – Twitter recently said it would close down the Vine website and community. The new version of the Vine app retains the 6.5 second video format, but the videos will be published directly on Twitter's feed and thus more closely integrated with its social network[4].

[1] https://techcrunch.com/2010/12/09/youtube-time-limit-2/

[2] http://www.theverge.com/2016/10/28/13456208/why-vine-died-twitter-shutdown

[3] http://www.theverge.com/tech/2016/3/29/11325294/instagram-video-60-seconds

[4] http://www.theverge.com/2017/1/5/14175670/vine-shutting-down-rebrand-download-archive

This paper aims to examine how crucial changes such as social network integration and time limit expansion might affect user engagement with this new format. To better understand these issues, we formulate the following research questions:

RQ1 What are the relative roles of social and content quality factors in driving engagement and popularity in micro-videos?

RQ2 How does the strict time limit impact video quality, and user engagement (both as creators and consumers) with such videos?

We answer these questions from an empirical perspective, using a dataset of nearly all ($\approx 120,000$) Vine videos that were uploaded to one of the 18 globally available channels on Vine during an 8 week period. We complement these with other datasets including a curated dataset (*POP12K*) of 12,000 popular Vine videos, as well as samples from other micro-video platforms such as Instagram[5].

To address **RQ1**, we take the three metrics of popularity we collect – counts of loops, reposts and likes – as quantification of the *collective* user engagement of the consumers of a video, and ask to what extent the content- and social network-related features affect these metrics. To answer this question, we adopt a novel methodology. We train a random forest classifier that, given a threshold for a metric of popularity, is able to able to distinguish items on either side of the threshold into popular and unpopular classes with high accuracy, precision and recall, using the features we have identified. The relative importance of different features then gives an indication of the extent to which those features affect the metric under consideration. We progressively consider higher and higher threshold values for videos to qualify as popular or engaging, and thereby identify trends and changes of relative importance of different features. Interestingly, we find that as the threshold for popularity becomes more and more stringent, features that represent quality of the content become collectively as important as social features such as the number of followers. Echoing an effect also observed in Instagram photos [3], we find that presence of faces in vine videos significantly increases engagement, and is the most important content-related factor.

Next, to explore **RQ2** we look at how the quality of the video varies over time in micro-videos, and discover a *primacy of the first second* phenomenon: the best or most salient parts of the video, whether in the aesthetic space or affect (sentiment) space, are more prevalent in the initial seconds of the micro video, suggesting that the authors are consciously or subconsciously treating micro-videos similar to images – in the initial part, the video is composed with aesthetics and affective quality in mind, resulting in a higher quality level;

[5]On acceptance, our newly crawled data will also be shared for non commercial research

but quality declines as the video plays over time. Furthermore, echoing the primacy of the first second phenomenon, we find that the quality of the first seconds of the video are as effective as the quality of the whole video in predicting popularity/engagement. Fig. 2.1 shows examples of these effects through two videos in our dataset of popular videos. In both videos, we observe content quality deteriorate over time, illustrating the primacy of the first seconds.

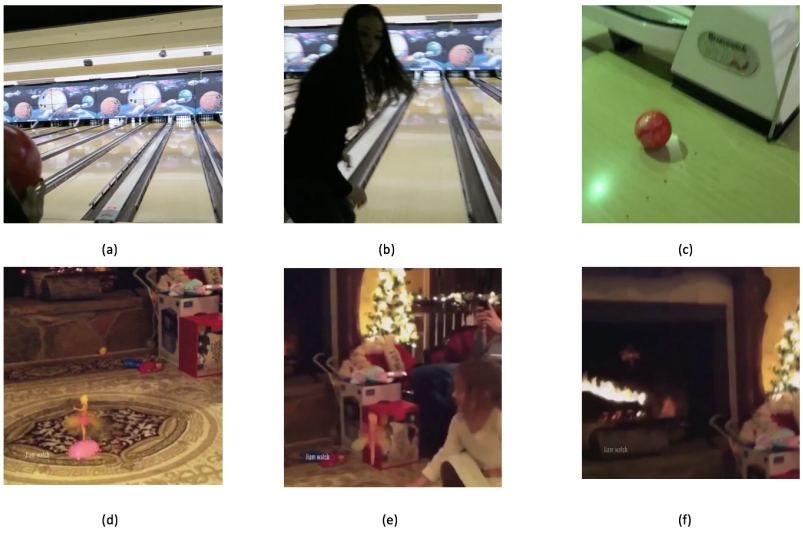


Fig. 2.1 *Vine Samples from first, second and thirds one thirds of the video. Images (a) , (b) and (c) show a progressive drop in brightness and sharpness due to shaky camera. Images (d) (e) and (f) shows a progressive drop in contrast.*

We confirm these computationally acquired findings with real user impressions by designing a survey which was answered by 115 respondents: Over 66% of users react to (like/comment on) content from their friends, making social interaction a significant part of content consumption; and 44% of users form opinions about videos in the first few seconds, validating the observed primacy of first seconds effect. Our survey also suggests that platforms such as Vine are seen as less professional and more immediate formats than, say Flickr images or YouTube videos, providing support to David Pogue's position that micro videos are a new kind of user-generated content [57], and therefore should be treated differently when it comes to user engagement.

2.2 Related Work

Our paper closely relates to those works in machine vision that infer intangible properties of images and videos. While computer vision frameworks typically focus on analysing image

semantics using deep neural networks [35], researchers have started exploring concepts beyond semantics, such as image memorability [27], emotions [43], and, more broadly, pictorial aesthetics [10, 18, 42]. This work specifically focuses on on-line visual content collected from social media. Researchers have shown that, by leveraging social media data in combination with vision techniques, systems can estimate visual creativity [60], sentiment [30, 78] and sarcasm [65].

More specifically, our work closely relates to research that combines social media studies and computer vision to analyse popularity and diffusion for social media posts: for example, Zhong *et al.* were able to predict the number of post "re-pins" given the visual preferences of a Pinterest user [84]; recent work [44] has also used multimodal features to predict the popularity of brand-related social media posts. Different from these works which focus on prediction, this paper looks at understanding user engagement.

Media popularity prediction studies generally focus on non-visual features. For example, [79] used textual annotations to predict various popularity metrics of social photos. Social metrics such as early views [56] or latent social factors [51] have also been used to effectively estimate video popularity. However, the fact that many popular media items may not depend on the social network [7] suggests that intrinsic media quality is an important factor for diffusion, engagement and popularity, which we explore in this paper.

Recent work in the field has explored the importance of visual content in analysing popularity: [75] analysed the visual attributes impacting image diffusion, and [64] studied relations between image quality and popularity in on-line photo sharing platforms. Bakhshi et al [3] showed that pictures with faces tend to be more popular than others. Similar to our paper, researchers have used computer vision techniques to estimate image popularity in Flickr [34]. Moreover, a work done by Fontanini et.al [20] explores the relevance of perceptual sentiments to popularity of a video. Unlike these works, we explore content features to fully understand user engagement and popularity in micro videos, a new form of expression radically different from both the photo medium and the video medium.

Micro videos are relatively new, so work specifically on micro video analysis has been limited. Redi *et al.* [60] quantify and build on the notion of creativity in micro-videos. A large dataset of 200K Vine videos was collected by Nguyen et.al [50], focusing on analysis of tags. Closest to our work is Chen *et al.* [8] who use multimodal features to predict popularity in micro videos. However, although we use popularity prediction as an intermediate tool, our focus is on understanding impact and importance of different features in determining popularity or engagement. To this end, we introduce a novel methodology that allows understanding up to which point social features are prominent over content features. Additionally, we demonstrate the "immediacy" of engagement with micro videos by showing

that the content from the first two seconds of the video is just as good at predicting popularity as the entire content . Collectively, these results allow us to characterise Vine as a new medium of expression, different from previous work.

2.3 Introduction to Datasets

Micro videos were pioneered and popularised by Vine[6], which was launched in 2012. Vine videos are constrained to a maximum length of 6.5 seconds. Videos are typically created using the mobile app and posted on user's profile which can be followed and shared by other users within the app or the website. We stress most of our work on videos sampled from Vine, complemented by Instagram data, which will be introduced as appropriate. The rest of this section gives details about the Vine datasets.

2.3.1 Dataset description

Dataset	Posts (total)	Loops/Views (median)	Reposts (median)	Likes (median)
POP12K	11448	318566	2173	7544
ALL120K	122327	80	0	2

Table 2.1 Summary characteristics of datasets used

The data used in this paper is summarised in Table 2.1, and was collected in two phases as described below:

Popular videos dataset First, we collected $\approx$ 12,000 videos which have been marked by Vine as 'popular', by tracking the 'popular-now' channel[7] over a three week period in Dec 2015, and downloading all videos and associated metadata once every six hours, and removing any overlapping videos from the previous visit. The crawling period was chosen to ensure that consecutive crawls have an overlap of several videos, and this sufficed for all visits made to the website during the data collection period; thus the dataset we collected is a complete collection of all 'popular-now' vines during the 21 days under consideration.

Vine does not disclose the algorithm used to mark a Vine as popular; yet we observe (see Table 2.1) orders of magnitude more loops, reposts and likes in the popular-now dataset than in the non-popular dataset. Thus we believe that the algorithm used by Vine to select vines

[6] http://vine.co
[7] https://vine.co/popular-now

for the 'popular-now' channel is strongly affected by the numbers of loops/revines/likes. Note that the numbers of loops etc. were collected at the time of crawl, within a maximum of six hours of being posted on the 'popular-now' channel, which limits the possibility that the counts increased *as a result* of being featured on the popular-now channel. In the rest of the paper, we use the counts in the popular-now dataset to calibrate the definition of 'high engagement'. While there is a possibility that this is a biased proxy for global engagement, it nevertheless provides a baseline against which to compare all videos.

All channel videos dataset In the second phase, we collected videos accessible from each of the 18 global Vine channels or categoriesover a period of 8 weeks from Aug 16 to Oct 12 2016. Again, a crawling period of six hours was chosen for consecutive visits to the same channel, and the 100 most recent vines were fetched with each visit. The number 100 was a result of an API limit from Vine. Our dataset captures nearly all videos uploaded to Vine and assigned to a channel. The only exception is the extremely popular comedy channel, for which we nearly always find more than 100 new videos (we only download the 100 most recent videos for the comedy channel). In total, this results in a dataset of ≈120,000 videos. We track loop, revine and like counts over time, periodically updating each video's counts every three days until the end of data collection. At the last tracking cycle, we have metadata for each post for 3 weeks after initial upload.

Note that while we obtain nearly all videos across the channels, our dataset does *not* capture *all* videos uploaded to Vine – Vine creators do not need to assign a video to a channel. However, due to the Vine platform structure, vines that are not in channels have near-zero probability to get seen by other users apart from the followers. We use channels to restrict ourselves to vines which have a chance to get exposed to a reasonably global audience of those interested in a topic category, and therefore to vines that have a higher potential for garnering high engagement.

2.3.2 Feature Descriptions

In order to fully understand how micro-videos engage users, we characterize the content of videos using computer vision and computational aesthetics techniques and extract a number of features (Table 2.2 in Appendix), which can be divided into the following categories:

Image quality features These features are mostly taken from computational aesthetics literature, and have been recognized as heuristics for good photography. Prior work [84] has identified a set of image quality features that robustly predict user interest in images. We adapt these to videos by computing the features on images taken at regular intervals from the video under consideration, and use the values to understand intrinsic quality of Vine videos.

We use a combination of low-level features such as contrast, colourfulness, hue saturation, L-R balance, brightness and sharp pixel proportion, together with higher level features such as simplicity, naturalness of the image, and adherence to the "rule of thirds" heuristic.

Audio features Following previous work on micro videos [60], we use audio features known to have an impact on emotion and reception. Using open source tools [37, 38], we measure *loudness* (overall volume of the sound track), the *mode* (major or minor key), *roughness* (dissonance in the sound track), and *rhythmical* features describing abrupt rhythmical changes in the audio signal.

Higher Level features Affect (emotions experienced) is well known to strongly impact on user engagement [39, 52]. To understand the sentiment conveyed by the video frames, we use the Multi Lingual Sentiment Ontology detectors [30] which express visual sentiment of video frames on a scale of 1 (negative) to 5 (positive). We sample frames at regular intervals and compute the affect evoked by these frames using this 5-point scale. Another higher level feature we consider is the presence of faces, which has previously been shown to have a strong influence on likes and comments in image-based social media [3]. We therefore adapt it to the video context by computing the *fraction of frames with faces*. Finally *Number of past posts* by the creator of the video under consideration is also included to reflect user experience and activity on the social media network.

Social features We consider the *number of followers* of the author of a content as a direct feature to reflect the user's social network capital.

A more detailed description of all the features can be seen in Table 2.2.

2.4 User Engagement in micro videos

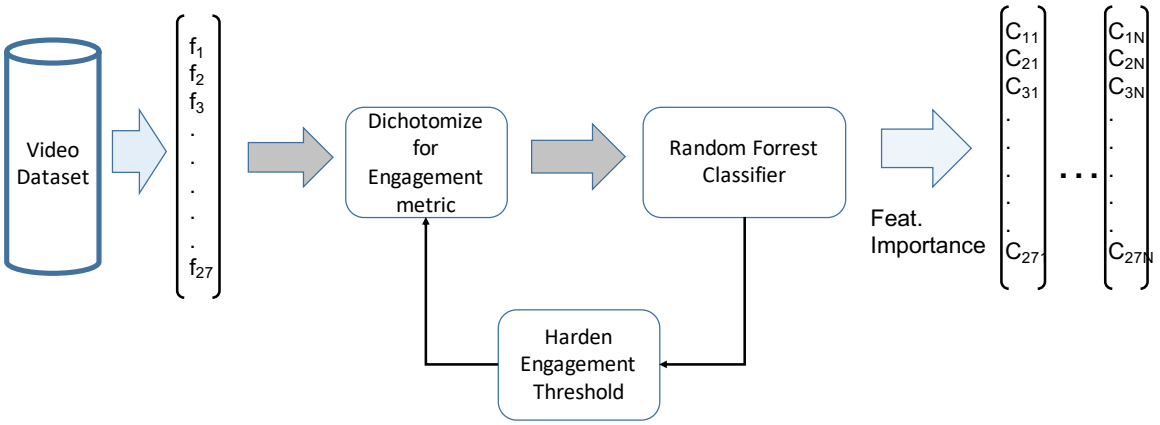
We begin our analysis by devising a novel methodology to analyze how the previously defined features impact user engagement in micro videos (**RQ1**). Our results indicate the importance of social features for highly engaging videos, and that the presence of faces is a strong content-related feature that positively impacts user engagement.

2.4.1 Metrics and methodology

To understand which aspects or features are important for user engagement, we need to: 1) define a metric for engagement, and 2) develop a methodology to study how the metric is influenced by different features.

Features	dim	Description
Visual Quality Features		
RMS contrast	1	RMS contrast is calculated as standard deviation across all the pixels relative to mean intensity
Weber Contrast	1	Weber contrast is calculated as $F_{weber} = \sum_{x=width} \sum_{y=height} \frac{I(x,y)-I_{average}}{I_{average}}$
Gray Contrast	1	Gray contrast is calculated in similar to RMS contrast in HSL colour space for the L value of pixels.
Simplicity	2	Simplicity of composition of a photograph is a distinguishable factor that directly correlates with professionalism of the creator [33]. We calculate Image simplicity by two methods: Yeh simplicity [81] and Luo simplicity [42].
Naturalness	1	How much does the image colors and objects match the real human perception?To compute image naturalness we convert the image into the HSV color space and then identify pixels corresponding to natural objects like skin, grass, sky, water etc. This is done by considering pixels which an average brightness V $\in$ [20 , 80] and saturation S > 0.1. The final naturalness score is calculated by finding the weighted average of all the groups of pixels. [84].
Colourfulness	1	A measure of colourfulness that describes the deviation from a pure gray image. It is calculated in RGB colour space as $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$ where $rg = R - G$ and $yb = \frac{R+G}{2}$ and μ and σ represent mean and standard deviation respectively
Hue Stats	2	Hue mean and variance which signifies the range of pure colours present in the image. It is directly derived from the HSL colour space
LR balance	1	Difference in intensity of pixels between two sections of an image is also a good measure of aesthetic quality. In non-ideal lighting conditions, images and videos tend to be over exposed in one part and correctly exposed in other. This is generally a sign of amateur creator. To capture this we compare the distribution of intensities of pixels in the left and right side of the image. The distance between the two distributions is measured using Chi-squared distance.
Rule of Thirds	1	This feature deals with compositional aspects of a photograph.This feature basically calculates if the object of interest is placed in one of the imaginary intersection of lines drawn at approximate one third of the horizontal and vertical positions. This is a well known aesthetic guideline for photographers.
ROI proportion	1	Measure of prominence given to salient objects. This measure detects the salient object in an image and then measures proportion of pixels its relative to the image
Image brightness	3	Features signify brightness of the image. Includes average brightness, saturation and saturation variance
Image Sharpness	1	A measure of the clarity and level of detail of an image. Sharpness can be determined as a function of its Laplacian normalized by the local average luminance in the surroundings of each pixel, i.e. $\sum_{x,y} \frac{L(x,y)}{\mu_{xy}}$, with $L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$ where $\mu_{x}y$ denotes the average luminance around pixel (x, y).
Sharp Pixel Proportion	1	Out of focus or blurry photographs are generally not considered aesthetically pleasing. In this feature we measure the proportion of sharp pixels compared to total pixels. We compute sharp pixels by converting the image in the frequency domain and then looking at the pixel corresponding to the regions of highest frequency [81], using the OpenIMAJ [23] tool.
Higher Level Features		
Face Percentage	1	Percentage of frames in a video, which have been tested positive for at-least one face. Faces detected using Viola Jones Detector [77]
Frame sentiment	1	Median frame sentiment of all the sampled frames from a micro video. The sentiment was calculated using the Multilingual Visual Sentiment Ontology detector [30]
Past post count	1	Number of past posts user has uploaded prior to current one. This is a good measure of user's experience with the platform and activity.
Audio Features		
Zero Crossing rate	1	Zero crossing rate measures the rhythmic component an audio track [38]. It ends up detecting percussion instruments like Drums in the track
Loudness	2	This feature expresses overall perceived loudness as two components. Overall energy and average short time energy [37]
Mode	1	This feature estimates the musical mode of the audio tract (major or minor). In western music theory, major modes give a perception of happiness and minor modes of sadness. [38]
Dissonance	1	Consonance and dissonance in an audio track has been shown to be relevant for emotional perception [38]. The values of dissonance are a calculate by measuring space between peaks in the frequency spectrum of the audio track. Consonant frequency peaks tend to be spaced evenly where as dissonant frequency peaks are not
Onset Rate	1	This measures the the Rhythmical perception. Onsets are peaks in the amplitude envelop of a track. Onset rate is measured by counting such events in a second. This typically gives a sense of speed to the track.
Social Features		
Followers	1	Number of followers that the user posting a video has. This is the prime social feature available from the user meta-data. The number of followers directly represent the audience which are highly probably to engage with the video on upload.

Table 2.2 Dimensionality and description of features used to describe Vine videos



Defining a metric for user engagement: In this paper, we use *number of loops* of a micro video as a proxy for user engagement towards it[8]. Although user engagement is a broadly used term, and other metrics may well be used to represent user engagement, our choice is in line with previous related social media studies (e.g. [3]) that have used social attention metrics such as likes and comments to study user engagement. Video hosting platforms like Youtube also use the number of views (similar to number of loops on Vine) as a core metric for their user engagement API[9]. In the rest of the paper, we will use popularity and engagement interchangeably.

Motivating the methodology: Given a set of features, if we can build a machine learning model that uses the features to predict which content items are highly engaging, the relative importance of the different features in making the prediction can tell us about the relationship between the features and engagement. However, our results will only be as 'good' as the model is in predicting loop counts. Since predicting popularity with exact numbers such as loop counts is a hard problem, we turn to a simpler one: We define an arbitrary threshold count for loops, and categorize micro videos as popular or unpopular depending on whether the loop count is over or under the threshold. We then design a classifier that predicts whether a micro videos is popular or unpopular (alternately, as engaging or not) based on our set of 28 features (Table 2.2) . As discussed next, a simple random forest classifier can be trained to make this prediction with high precision and accuracy. The relative importance of different features then tells us about how the features affects user engagement.

This method has one major limitation: its dependence on the arbitrarily defined loop count threshold. Therefore, we conduct a sensitivity analysis by training a series of binary

[8]We obtained similar trends using number of reposts, but only report results with loops. Note that the loop counts of videos are highly correlated with reposts and likes. For example for videos in POP12K, $corr(Loops, Likes) = 0.80$, $corr(Likes, Reposts) = 0.91$, $corr(Reposts, Loops) = 0.74$.

[9]https://developers.google.com/youtube/analytics/v1/dimsmets/mets

classifiers for different loop count thresholds. This also allows us to study shifts in relative importance, as we move up the scale towards more popular and engaging objects, by defining increasingly higher numbers of loop counts as the threshold for categorizing a video as popular (or engaging).

2.4.2 Model details

Setup We sample 12,000 videos from our dataset, out of which 6,000 are popular videos from POP12K, and 6,000 randomly sampled from the ALL120K dataset, thus representing the entire spectrum of engagement levels. In each video, we sample the video track for individual frames at every second, and extract the audio track as well as meta-data related to the video and its author. Using these, we then compute the 28 dimensional vector of all the features in Table 2.2 and train a random forest classifier to distinguish popular and unpopular videos for different thresholds of popularity. We used the implementation from the *SKLearn* package with $\sqrt{n_{features}}$ split and 500 estimators, which provided the best trade-off between speed and prediction performance.

Performance Results Different classifiers are trained using the above method for different engagement/popularity thresholds, using an 80-20 split for training and validation. Fig 2.2c shows how these perform as we vary the threshold of "engagement" (popularity) from 80 loops (the median for ALL120K) to $\approx$ 500,000 loops (1.5 times the median of the popular videos i.e., POP12K). At each training iteration with a changed "engagement" threshold, we re-balance the dataset by choosing equal number of samples which fall in either classes. We take care that we are training on at-least 20% of the complete dataset by the end of the process, and stop increasing the threshold beyond that point to avoid over-fitting. The classifiers gave consistently high performance on the validation dataset (see lines labeled 6 sec), never dropping below 90% for accuracy, and 80% F-1 score, validating our next results about the importance of different features.

2.4.3 Feature analysis and implications

The impact of individual features on user engagement is calculated using Gini importance [41], and combined into social- and content-related (i.e., audio and video-related) features as described before (§2.3.2). Fig. 2.2a shows the trends in feature importance as a function of engagement threshold used (see lines labeled 6 sec). We observe that at lower thresholds of popularity, social features are much more important than content-related features, but at higher thresholds, content-related features increase in importance to become just as important as social features, suggesting that *content quality is important for user*

engagement at the top end of engagement. This facet of users' engagement with Vine might legitimize Twitter's decision to more closely integrate the Vine platform with its social network: since a large part of micro-video popularity can be explained with social factors, a better social network might further foster engagement with this unique form of expression.

We drill down further in Fig. 2.2b, and examine the importance of different kinds of content-related features. For each class of content-related features, we plot the mean of the feature set of the class. We observe that in terms of effective importance of different feature tracks, sentiment is the weakest influencer in the classifier decision process. We conjecture that the relative lack of importance of sentiments may partly be due to the extremely short nature of micro videos, which does not let emotional 'story arcs' and plots (e.g., drama) to develop as strongly as in longer videos.

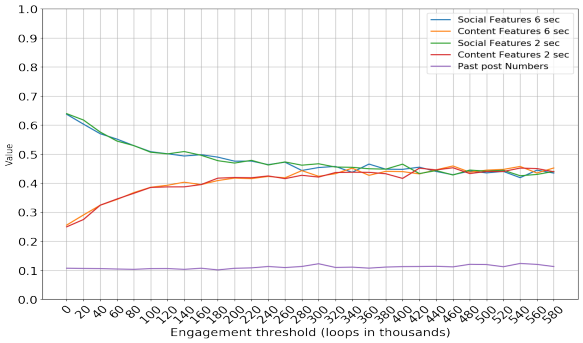
Further, we observe that the presence of faces in a frame strongly outweighs all other content-related features in predicting popularity. We confirm this in Figure 2.3 by comparing the percentage of faces in popular POP12K videos with the corresponding percentage in ALL120K videos (which contain a large number of unpopular videos as well as a few popular ones). These results indicate that popular videos tend to have more faces, i.e., "*faces engage us*". This is in alignment with similar results on other platforms, which also indicate that faces greatly enhance popularity related metrics such as likes and comments [3].

2.5 Primacy of the first seconds

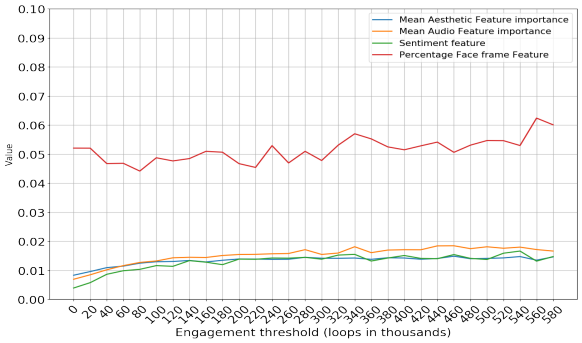
Next, we try to understand these findings further by examining the quality of the individual frames of the videos: One way to think about videos is as a sequence of images. With micro videos, this sequence is of course much shorter than in other videos, and we investigate whether this has impact on video quality (**RQ2**). Our results show a "primacy of the first seconds" effect, with quality deteriorating over time and the quality at the beginning is as good a predictor of engagement as quality of the entire video.

2.5.1 Image quality deteriorates over time

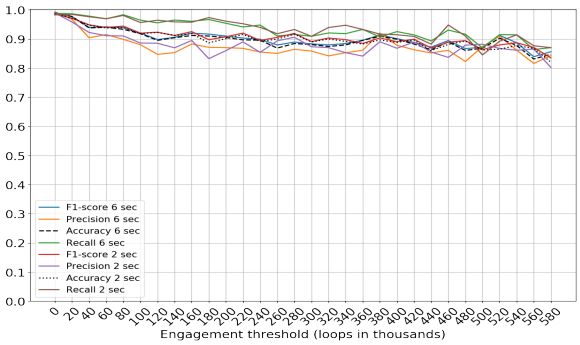
Vine videos can be at most 6.5 seconds long. We sample the videos twice every second and represent the whole video as a series of 12-13 static frames. This sampling rate is not too low to miss any considerable frame transitions, neither is it too high to include a lot of mid transition frames. For each sampled frame, we calculate the feature under consideration – sentiment, percentage of faces, and aesthetic score. To compute the aesthetic score, we extract the 18 aesthetic features described in Table 2.2. for each frame frame. To find an



(a) Content Vs Social Features



(b) Individual Content Features



(c) Classifier Performance

Fig. 2.2 Understanding engagement for different thresholds (min. number of loops considered as engaging). Two different classifiers are used, one using quality of the entire micro video (labeled 6 sec), the second measuring quality from only the first two seconds (labeled 2 sec). (a) As threshold becomes higher, content-related factors become as important as social factors (both classifiers). Note that unlike content quality computed from the first 2 seconds ('Content features 2 sec') rather than the entire 6 seconds of the video ('Content features 6 sec'), 'social features 6 sec' uses the same feature values as Social Features 2 sec', but the two are plotted separately to show the relative importance of social features in the 6 second vs 2 second classifier. (b) Amongst content features alone, presence of faces in the video is the single most dominant feature, across all threshold levels (6 second classifier) (c) Both 2 sec and 6 sec classifiers perform similarly across all metrics such as Precision, Recall and F1-score. Performance is high across all engagement thresholds: all metrics are consistently over 0.8 or 0.9.

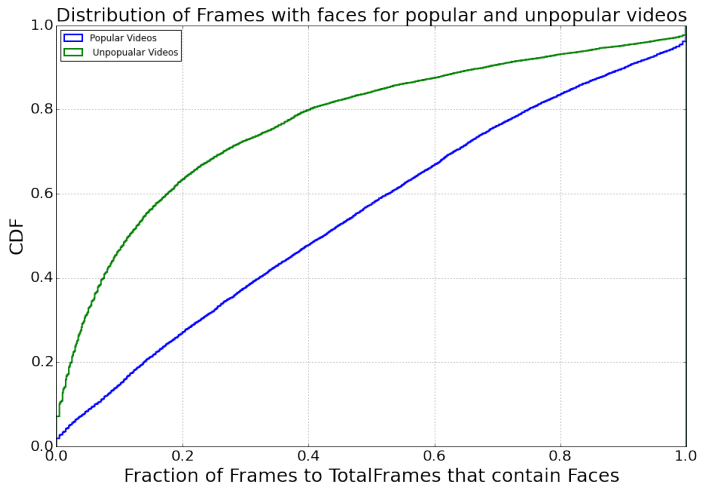


Fig. 2.3 *CDF for popular and unpopular videos. The CDF signifies the cumulative distribution of percentages of frames containing faces in a vine video. The observation here is popular videos tend to have higher face percentage than unpopular videos*

aggregate overall aesthetic score of each frame, we use a weighted sum of all the features (This is possible because all the features are on the same scale), where the weights are calculated to be proportional to the importance of each feature in the classifier designed in the previous section.

For each video and each feature, we then compute when in the video the feature reached its maximum value. We then divide the videos into two second intervals, essentially dividing the video into its first third, second third and third third. We then ask what proportion of videos had the maximum value of a feature in the first (respectively second and third) third. This procedure tells us when we are likely to find the 'best' part of the video.

Fig. 2.4 shows the result for each major category of content-related feature, plotted over both our datasets (ALL120K and POP12K). We observe a general trend where the first third has the maximum (best) value for all features considered. For instance, the best aesthetic score is to be found in the first two seconds. Similarly, the proportion of faces, an important predictor of engagement (Fig. 2.3), is also maximum in the first third.

Note that for sentiment values, the minimum value is just as valid and valuable as the maximum, representing a sad or emotionally dark segment of the movie with negative sentiment, in contrast to a happy segment of the movie with positive sentiment. Therefore, we calculate which third of the movie we find the maximum and minimum sentiment values and plot these separately. In both cases, we find yet again that the first third of the video has the maximum (minimum) sentiment value for the majority of videos.

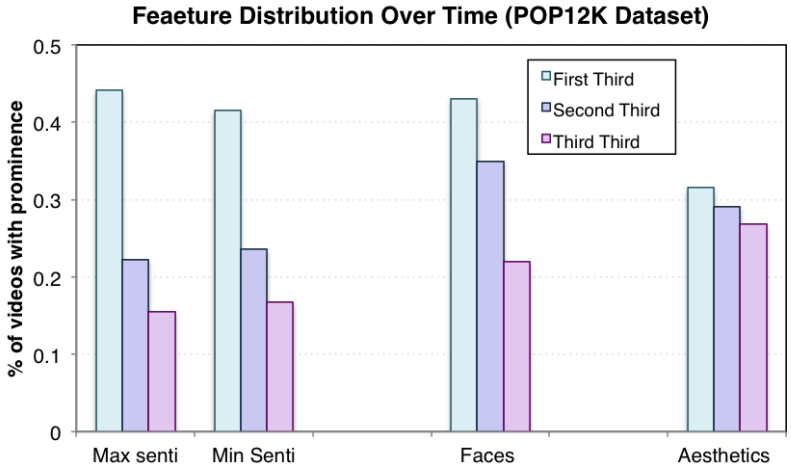


Fig. 2.4 *Evolution of Feature magnitude: The graph shows sharp trend in prevalence of strongest component of a feature in the first one third of the video. The strength decreases progressively for the successive thirds. (Results shown for POP12K dataset. Similar results obtained for ALL120K.)*

2.5.2 Loops and likes are obtained on first sight: Initial seconds predict engagement

Collectively, the results above paint a picture where the first seconds of the micro video are highly important in engaging the user. We conjecture that this might be because of the mobile-first nature of Vine: the primary user interface is the Vine app, where users select which videos to watch by scrolling over it. The vine only plays when the user retains focus over the video, and hence the first seconds are likely critical for grabbing user attention and engaging them.

We next take this result to its logical conclusion, and ask how the classifier developed in the previous section for predicting engagement would work if using only content-related features from the first third of the video rather than from the whole micro video. Following the same methodology as described in the previous sub-section, we develop a series of classifiers for different popularity thresholds, training this time on image content-related features drawn from the first two seconds of the video rather than from across the whole video. The same set of hyper parameters were used as in the previous setting. As shown in Fig. 2.2c, the resulting classifiers (labeled 2 sec) perform very similarly to the classifiers developed before (labeled 6 sec). Further, Fig. 2.2a shows that the relative importance of different features is also nearly identical to the previous results. It should be emphasized that

although these results were obtained using loop counts as the metric for user engagement, similar results have also been obtained using reposts (re-vines).

These results point to a *primacy of the first seconds* effect, whereby the first seconds of a micro video matter as much as the whole video, suggesting that they behave almost like still images in terms of user engagement.

2.6 User study

To complement the data analysis and gain deeper insight into what drives user actions and engagement, we designed an anonymous user study which captures user behavior when engaging with micro-videos.

2.6.1 Survey methodology

We initially recruited undergraduate students, obtaining about 33 responses. Subsequently, we tweeted the survey out to the Official Vine Twitter account and to the accounts of Vine and Instagram users, in order to gain further exposure amongst users of these platforms. In all, 115 users responded to our survey. Table 2.3 summarizes the respondents' demography and usage preferences. Most questions asked were to be answered either on a 5-point Likert scale (Strongly agree to Strongly Disagree), and or in a semantic differential format, with three options to choose from.

Attribute	**Value**
Male respondents (%)	44.2
Female respondents (%)	55.8
Age Demography (%)	
18-24	43.4
25-31	34.5
32-40	14.2
40+	8

Table 2.3 Summary of survey responses

2.6.2 Validation of data-driven results

To understand engagement with micro-videos and validate the findings that emerged from our analysis of **RQ1** and **RQ2**, the survey asked the following 5 questions to be answered on a 5 level Likert scale (strongly agree - strongly disagree):

A I tend to like/comment on videos from friends rather than from strangers

B I always form an opinion of a video in the initial few seconds, once the video starts playing

C I rarely watch short videos (Snapchats, stories) , all the way to the end.

D I prefer to watch short videos of humans on these platforms. E.g. I like to see a person talking/expressing rather than outdoor scenery, or Cats.

Almost 66% users agreed to question **A**, which reaffirms the tendency of socially embedded users being able to get high engagement scores. 44% of users agreed to question **B** (and further $\approx$ 30% users were neutral) and 38% users agreed with statement **C**, supporting the observed the "primacy of the first seconds" effect. Contrary to our findings regarding faces, only 34% users agreed with statement **D** (although a further 39% remained neutral; thus only a minority 27% of users disagreed or disagreed strongly). Such result might suggest that for many users, our attraction towards face shapes is innate [69] and people do not consciously engage more with faces.

2.6.3 Understanding what matters to users

The next part of the survey went beyond confirming the data-driven analysis by asking the respondents how their behavior changed when it comes to *acting* on a video they engaged with, i.e., when do they like/forward, comment or stop playing the video? 44% like, comment or share videos only after finishing watching it, and but a sizable 56% agreed that they do so in the middle of watching the video itself, or right at the beginning (19% share at the beginning. 37% somewhere in the middle); again pointing to the need for capturing users in the initial parts of the video.

Interestingly, a majority of 55% of respondents agreed (on a 5 point Likert scale) to the statement: "I don't really care about the quality of the micro-video or stories, as long as I like the content". This result, together with the previous answers seems to imply that the fall off in content quality in the latter parts of micro videos does not negatively impact user engagement. However, users do see a difference between micro videos and "traditional" (and older) user generated content platforms such as YouTube: an overwhelming 75% of respondents rate the production quality of YouTube videos quality as more professional than micro videos.

2.7 Discussion and conclusions

In this paper, we took a first look at user engagement with micro videos. Defining engagement in terms of social attention metrics such as likes, revines (reposts) and loop counts, we find that content quality-related features have as strong an influence as social network-based exposure in driving these metrics. Furthermore, the quality of the first couple of seconds is higher than the quality of subsequent seconds, and can predict whether a micro video will be engaging or not, just as well as looking at the quality of the entire video. We further conduct a user study to understand ground-truth user behavior when it comes to micro-videos. The study suggests that users tend to make quick opinions regarding micro-videos and engage with them almost in an image-like fashion, where they may begin but not finish watching the short 5-10 second long video.

These aspects of micro video user engagement has important implications and bearing on future work:

1. Advertisements on the Web are driven by social attention metrics. Therefore advertisers need to know and adjust their strategies based on the insight that user attention is driven to a large extent by the initial seconds. Although video ads do not appear to be common in today's micro videos, how to place ads that grab user attention within a short duration of time will be a problem that is interesting both from a research and a business perspective.

2. A possible reason for the deterioration of image quality is that it may be difficult to maintain image composition, focus etc using a mobile phone camera with moving subjects. Novel UI and multimedia techniques that can help correct for such quality deterioration could greatly help micro video creators – and also represent a second promising direction for further study.

3. Recently, several micro video platforms have started extending the duration of micro videos. Although the wisdom of longer micro videos without appropriate editing tools has been questioned[10], from a research perspective it would be interesting to study how user behavior and engagement changes as longer micro videos become more common place. Interestingly, we find that in a small sample of about 6000 Instagram videos (where the maximum permitted duration is 60 seconds), users continue to prefer shorter videos, with 70% of videos less than 20 seconds long, and the median duration at just under 15 seconds. Such user preferences can and should be considered as the micro video format evolves further on different platforms.

[10] http://www.theverge.com/2013/6/20/4448906/video-on-instagram-hands-on-photos-and-video

More generally, in this work we considered user engagement as a single dimension. However, we acknowledge that user engagement is a very subjective notion, impacted by different factors including user location, habits, gender, visual preferences. In future work, we plan to explore how such different user sub-cultures perceive and engage with micro videos, following recent works from the Multimedia community studying the impact of culture in subjective image perception [30]. A second dimension to explore in our future work is generalising the above findings to other micro video platforms – our preliminary studies indicate that key results such as the Primacy of the first seconds effect, are robust across platforms, applying to Instagram videos as well. However, more work is required in this direction.

CHAPTER 3

ONLINE HEALTH FORUMS PRIMER

Online communities have the potential to influence health and health care. Recent studies have suggested that the participation of people with long-term conditions (LTCs) in online communities (1) improves illness self-management [1], (2) produces positive health-related outcomes [2-4], (3) facilitates shared decision-making with health care professionals [5,6], and (4) may even reduce mortality [7].

There is also evidence that self-management support interventions can reduce health service utilization [8,9].

Online communities have experienced an upsurge in popularity among people with chronic respiratory conditions such as cystic fibrosis [10], asthma [11], pulmonary hypertension [12] and chronic obstructive pulmonary disease (COPD) [13]. More than 15 million people in England suffer from a long-term condition or disability, and they account for at least 50 percent of all general practitioner appointments [14,15]. Thus, assessing how these online communities function and evolve can have important implications for health care provision.

This form of "user-led self-management" of LTCs bears similarities with the "expert patient" model, an approach to self-management of LTCs produced by the United Kingdom (UK) Department of Health in 2001 [16]. Evidence of the effectiveness of conventional off-line self-management programs based on the expert patient model, though, has been weak [17]. Clinic-based self-management programs often failed because of: (1) lack of awareness and engagement among patients and staff, (2) failure to consider low health literacy or cultural norms, (3) lack of attention to the need for family and social support, and (4) a fragmented approach to the provision of health and social care [18]. Although online health communities can be seen as an extension of the expert patient model, network effects, in addition to the online disinhibition effect [19], make them a distinct and unique complex intervention mechanism.

On average, one in four people with an LTC who use the Internet tries to engage online with others with similar health-related concerns [20]. In particular, it has been suggested that the value of participating in an online community lies in the possibility of gaining access to a range of people and resources quickly, easily [21], and anonymously [4], as well as obtaining tailored information and emotional support [1,22-26]. However, most of this evidence comes from qualitative studies [1,27], whereas only recent years have witnessed an increasing interest in quantitative assessments of online communities as intervention mechanisms [28-33]. Recent studies have been concerned with the users' unequal contributions and engagement patterns, and with the role of superusers. However, the contribution of superusers to the sustainability of online health communities and their structural properties remains mostly unclear.

The potential future integration of online health support systems with formal health care provision should be underpinned by a better understanding of how they are used and by evidence of their effectiveness. Indeed, as suggested by the Medical Research Council [34], integrating online support systems with the more traditional health care provision would require the identification and comparative assessment of potential alternative intervention mechanisms.

An expanding body of literature concerned with social network analysis has examined the structural patterns of relations among interacting actors and the social mechanisms that enable them to gain access to valuable resources [35]. There is also increasing evidence that network approaches can be applied to understanding the users' "expertise" [36], their interactions, and network effects on health-related outcomes in online health communities [37,38]. Uncovering the mechanisms underlying the formation of successful social networks requires a study of how online connections among people, namely the social ties or links, emerge and evolve, and how groups of individuals gradually grow in membership and become interconnected with one another. These processes of tie creation and group formation in online patients' communities are still mostly unexplored [1].

In this study, we performed a network analysis of the structure and dynamics of two online communities of people with LTCs. We chose the Asthma UK and the British Lung Foundation (BLF) communities as an exemplar of such communities because their users typically suffer from chronic respiratory conditions. In particular, while Asthma UK users typically suffer from a respiratory condition characterized by variable and recurring symptoms, BLF users represent a more heterogeneous population of participants affected by different diseases linked to chronic symptoms of breathlessness (eg, COPD, pulmonary fibrosis, cystic fibrosis, and lung cancer). We aimed to uncover and understand how these communities function and evolve, and the role that some users have in maintaining integration and cohesion (see

Textbox 1 for research questions). Ultimately, this study provides evidence for gauging the effectiveness of different interaction patterns and the users' structural positions and their potential for enhancing and sustaining health online communities as scalable self-management support interventions.

3.1 Dataset and properties

Data were collected by HealthUnlocked [39], the online platform provider of the Asthma UK and BLF communities. Registered users can choose to either write posts publicly or send private posts to one another. In the latter case, posts are shared between 2 users only, whereas when posts are written publicly, a large number of users can become connected through threads of posts. Only posts that were shared publicly were collected and analyzed. For this study, user identifiers (IDs) were anonymized by HealthUnlocked, and no demographic information was collected. The data sets included posts and their metadata (ie, the anonymized user ID numbers), user roles (eg, user, administrator, or moderator), date of posting, the hierarchical level of the post within the corresponding thread, and the dates in which the users joined and left the community. Both communities were moderated, and HealthUnlocked moderators (identified through metadata linked to posts) were included in the analysis to assess their contribution and compare it with other users. Online communities on the HealthUnlocked platform benefit from additional functionalities compared to other online forums, such as built-in patient groups that moderate the content. In particular, the content accessed by users is tailored to their interests, and profiles highlight users' condition, chosen community, medications and treatments they use or find interesting. No data were collected on participants' characteristics, though only people declaring themselves to be older than 16 years were permitted to create an account and take part in the online communities.

3.2 Graphs: A primer

3.3 Activity patterns of users

We looked at the number of users, the number of posts and connections per user and posting frequency. A connection (ie, a tie, link, or edge) was established from one user to another when the former replied to a post by the latter (see Textbox 2 for network analysis terminology). The pattern of connections generated over time through the cumulative number of posts and replies was examined. We were interested not just in the number of posts and responses but in who responded to whom, and when. To this end, we used social network

analysis [40] to visualize and study the structure of the relationships between users. Both visualization and analysis were conducted using the Gephi software. The network analysis was carried out through additional custom computer code in python. Descriptive analysis of the networks (ie, number of users, posts, and posting frequency) were calculated using the Pandas library, an open source library providing data structures and analysis tools for the Python programming language.

As a result of the small percentage of users who wrote posts to a disproportionally high number of users, the users' activity showed long-tailed distributions. Therefore, our analysis was based not only on means and standard deviations but also on medians.

To uncover time patterns in posting activity, we used Fourier transforms of the time series of the users' activity [46], a known method used for the analysis of signals. Through Fourier transforms, we identified the frequency components, called harmonics, that together made up the posting activity stream. In other words, we regarded the posting activity over the entire observation period in both communities as a complex signal and identified the frequency components that made up such a signal. This analysis was performed using custom code in Scipy, a Python-based scientific computing library.

The "rich-club" coefficient is a metric designed to measure the extent to which well-connected users tend to connect with one another to a higher degree than expected by chance [43]. To this end, for each value k of a node's degree (ie, the number of other users a given user is connected with), we computed the ratio between the number of actual connections between nodes with degree k or larger and the total possible number of such connections [47]. We then divided this ratio by the one obtained on a corresponding random network with the same number of nodes and degree distribution (ie, the probability distribution of the degrees over the whole network) as the real network, but in which links were randomly reshuffled between nodes. Thus, the rich-club coefficients may take values lower or higher than 1, depending on whether the real network has a higher or lower tendency to coalesce into rich clubs than randomly expected. In particular, networks that display a high rich-club coefficient (ie, greater than 1) are also said to show a "rich-club effect," namely the tendency to organise into a hierarchical structure in which highly connected nodes preferentially create tightly knit groups with one another, thus generating exclusive clubs of (topologically) rich nodes, as illustrated in previous work [48].

In our study, superusers were defined according to their cumulative activity over the entire observation period. In total, we identified 400 superusers. To uncover how many superusers were active within each week, we detected how many unique users, among the 400 identified over the entire period, were active within that time window.

Following Zhang et al [36], the "z-score" was used as a proxy for users' expertise. According to this measure, replying to many questions suggests one's expertise, while asking questions indicates lack of expertise. In our analysis, we treated anyone starting a thread as a help-seeker, and anyone commenting on the thread as a help-giver [36]. Accordingly, the proposed z-score aims to capture the combined help-seeking and help-giving patterns. To this end, for each user, we measured how many standard deviations the observed total number of the user's help-giving posts lies above or below the expected number of help-giving posts for the whole system. We extended the approach proposed by Zhang et al by empirically assessing the probability of posting and answering a question across all users over the entire observation period. In the BLF community, we found that the probability of answering is Pa=2/3, while the probability of posting is Pq=1/3. We assumed a Bernoulli process of posting an answer or a question to the forum, with probabilities defined as above. The z-score for a given user i was calculated according to equation (a) in Figure 1, where ai refers to the total number of answers user i posted to the forum, qi is the total number of questions user i asked in the forum, and ni=ai +qi is the total number of messages posted by user i. To obtain $Zscore_i$, let us define a random user that posts the same total number of messages nrandom to the forum as user i (ie, nrandom=ni). We would expect this random user to post an average number of answers to the forum given by equation (b). Plugging in the value of $Pa = 2/3$, we obtained equation (c). Similarly, we would expect the random user to post answers with a standard deviation given by equation (d). Plugging in the value of $Pa = 2/3$, we obtained equation (e). To measure how many standard deviations above or below the expected random value a user i lies, we then computed Zscorei according to equation (f). Plugging in the values of μ_{random} and σ_{random}, we obtained equation (g). Finally, by substituting $ni = ai + qi$, we obtained equation (h).

3.4 Dataset charactarization

The data sets span, respectively, 10 years for the Asthma UK and 4 years for the BLF communities (see Table 1).

Despite the shorter time span, as a result of the larger number of users, the number of posts in the BLF community was higher than in Asthma UK, namely 875,151 compared to 32,780 respectively. Moreover, BLF users wrote a higher number of posts per user and were connected with a higher number of other users when compared with people in the Asthma UK forum (see Figure 2). In both communities, 60%-70% of registered users wrote no posts (ie, they were lurkers). Users who wrote more than one post contributed with a median

of 8 (range 2-8947) and 5 (range 2-1068) posts in the BLF and Asthma UK communities, respectively.

The number of official moderators among the highly active users was negligible; there were no moderators in the top 5% contributors to BLF and only 2 in the top 5% for Asthma UK. Thus, our network analysis predominantly reflects content originated from registered users.

When classified according to posting activity (ie, number of posts written to the forum), the top 5% users contributed to a substantial proportion of all posts: 58% and 79% in the Asthma UK and BLF communities, respectively. Superusers were those who made a high number of connections with other users in both Asthma UK and BLF communities (see nodes of large size in Figure 2). Asthma UK superusers made a lower number of connections than BLF ones. The posting activity of these superusers will be analyzed in more detail in subsequent sections.

3.4.1 Activity

Posting Activity

The cumulative number of messages posted grew uniformly over time in the BLF community. By contrast, in 2015, the Asthma UK forum witnessed a substantial increase in posting activity, at a time coinciding with its move to the HealthUnlocked platform (see Figure 3A and B). This increase in activity can be attributed to the online community functionalities offered by HealthUnlocked, as described in the Methods.

The number of posts per user per week oscillated around a decreasing and an increasing trend (Figure 2C and D), while at the same time the number of posts always went up over the study period (Figure 1A and B). This suggests that there were intervals of time during which the rate of increase in new users was larger than the rate of increase in total posts. Moreover, in the Asthma UK forum users wrote according to two time patterns—they posted at an interval of 1-20 days or 6 months (Figure 4A), while those in the BLF community at an interval of 2 days (Figure 4B).

As more users joined the communities and connected to one another through online posts, distinct groups of connected users started to emerge. These groups, called network components (see Textbox 2), have fundamental implications for the effectiveness of processes of network dynamics such as information diffusion [49]. In a relatively short period, both communities underwent the formation of the “largest component” of connected users, namely a connected subset of users whose size increasingly outgrew the size of all other components (see Figures 1 and 4, and Multimedia Appendices 1 and 2). The largest connected components in both communities included 60

Figure 5 suggests that, as time went by, the number of forum participants and their posting activity increased, and the proportion of users who were part of the largest components decreased. This finding was expected because the number of posts also rose exponentially, yet at times at a lower rate than the one at which new users joined the communities (see Figure 1C and D). It, therefore, became more difficult for the network to self-organize into a connected component that would include 100% of the users. Figure 5A also shows that around week 450, when the forum moved to the HealthUnlocked platform, a larger fraction of users began to join the largest connected component, thus highlighting the role that the new online platform played in strengthening the connectedness of the network (see also Figure 3A and B).

3.5 Propensity to help

3.6 Not like conventional networks: Anti-rich club effect

3.7 Key takeaways, possible interventions

CHAPTER 4

STRUCTURE OF ONLINE SUPPORTIVE CONVERSATIONS

4.1 Introduction

The world has become more connected over the past decades thanks to the networked nature of the technologies of the day. It is seldom possible to spend a whole day without single interaction on the Internet. The Internet gives platforms where we can not only connect with our social counterparts but also exchange ideas and express opinions. These new mediums have become so ubiquitous, that some research suggests that they might be affecting our broader psychological state [6]. But on the positive side, studies have also proposed different ways in which this medium could be used for measuring and intervening in the matters of mental health[13, 14]. Online communities, or fora, offer a platform for users to directly interact with each other. Reddit[1] is one of the largest online communities which contains a number of sub-communities (so called *subreddits*) that can be about almost anything. On this platform, several subreddits are specifically tailored to mental health-related topics, such as *depression*, *anxiety* or *alcoholism*. These fora offer a unique opportunity to study the way people describe or discuss their problems in their own voice.

One of the most challenging, and devastating, global mental health concerns is suicide. Suicidal behaviour includes any thoughts, plans or acts someone makes towards ending their life. In health care services, preventing death by suicide is a priority, but accurately predicting whether or not someone is at risk of committing suicide is difficult. Moreover, a large proportion of deaths by suicide occur in populations that have never been seen by health service providers.you could probably improve the section above...

Several online platforms are used for expressing suicidal thoughts and reaching out for support. On Reddit, the subreddit *SuicideWatch* currently[2] has almost 94k subscribers, and is

[1] http://reddit.com/
[2] As of 27th June 2018

a moderated forum that is intended to offer peer support for people at risk of, or are worried about others', suicidal behaviour. The moderators take the message of peer support seriously, and are governed by guidelines that prohibits false promises, abuse, tough love and other clinically frowned upon methods of conversations[3]

As such it is valuable to understand what characteristics supportive communities like SuicideWatch have, and in what aspects such communities are similar or dis-similar from other casual subreddit conversations.

Recent studies have shown promising results in modeling and measuring signals and patterns in reddit communities related to mental health. For instance, statistical relations of mental health and depression communities with suicidal ideation have been studied [13, 14]. The authors explored linguistic and social characteristics that evaluate user's propensity to suicidal ideation. Approaches to classify reddit posts as related to certain mental health conditions have also been successfully developed, showing that there are certain characteristics specific to mental health-related topics in posts that can be automatically captured[21]. Furthermore, in a study focused on reddit posts related to anxiety, depression and post-traumatic stress disorder, the authors show that these online communities exhibit themes of supportive nature, e.g. gratitude for receiving emotional support[54]. Positive effects in participation in such fora have also been shown by improvements in members' written communication[53]. The supportive nature of comments in the SuicideWatch forum has also been studied by automatic identification and classification of helpful comments with promising results[32].

Most previous studies have aimed at studying the *content* of posts and their characteristics in relation to other posts. One important aspect of online communities is its supportive *function* — users turn to these platforms not only to express their thoughts and concerns, but also to receive support from the community. More references to be added in the introduction. Also, we need to add something about the Online disinhibition effect somehow.

To our knowledge, there are no studies that have specifically focused on modeling the supportive *nature* of online fora related to mental health. This work takes a macroscopic perspective, to quantitatively characterise and model the nature of supportive conversations. SuicideWatch is particularly interesting because of its purpose to offer peer support to people with suicidal thoughts, and also because of the complexity of this clinical construct.

Our aims in this study are to

- Understand similarities and differences between a Suicide watch conversation and a generic conversation using these abstraction.

[3]http://www.bbc.co.uk/newsbeat/article/35577626/social-media-and-suicide-what-its-like-being-a-moderato

Terminology	stands for
RP	Root post which begins a new thread on a subreddit
OP	Original poster who posts the Root post for a thread
SW	The suicide watch Subreddit
FP	Front page of Reddit.

Table 4.1 Notations and Terms.

- Study global properties of these conversations in comparison with control conversations.
- User network metrics to reason about global differences in terms of local interactions between users.

To model the network topology in an online community, we represent conversations in a forum using graph-based abstractions (users and replies) as described in Section 4.3.2. To measure global structure of these conversations, we user network topological metrics such as centrality: which measures importance of nodes in a network in terms of relaying information, branching factor: which measures how a conversation fans our over time, return distance: which measures how soon do users return back to the conversation and symmetric edges: which measures reciprocity of users in a conversation. To measure measure local interactions, we measure inter response times: which measure urgency of response to a message, semantic alignment between messages and local interaction motifs known as Triadic motifs : which gives an idea about how distinctive are interactions between subgroups of users.

4.2 Results

Particularly Suicide watch community consists of over 78k subscribers and reader, however is supported by mere 12 Moderators according to the latest count. The moderators are mainly present to prevent any kind of abuse, trolling or non-clinical or non-productive advices. These moderators do not have any form of formal training. However through several accounts they have confessed to learn through interactions and mentorship from more senior moderators on the site[4] All the moderators have been in that role for at least 3 years and the oldest goes as far as December 2008.

Our study is based on all conversations on SuicideWatch. We represent these conversations through networked abstractions as described in 4.3.2.

Through our analysis we find several discriminatory factors among Suicide watch conversations and generic front page conversation. We show that some of these factors are

[4]https://bbc.in/24rJYQH

predictive of suicide watch conversations to a very high degree. We also show that certain properties of these conversations can be backed by sociological theories of real life support conversations.

4.2.1 Peculiarity of threads of Support

We begin by characterizing the two networked abstractions, namely Reply Graphs and Interaction graphs as described in Section 4.3.2. We do so by first comparing these two abstractions with a baseline control conversation threads using certain macroscopic network properties. We first compare the number of unique users per thread. The two communities exhibit considerable difference. SW Sub-Reddit has a median of 5 users per thread and a mean of 6.7 users and BL threads have a median of 25 users and mean of 50 unique users. So this off the bat indicates that Suicide watch conversations are more intimate and involve less participants. Authors tend to participate on a similar level, with a median participation per author of Suicide Watch to be 5 and for baseline conversations to be 3.

How urgent are user responses?

Understanding the inter message times can act as a good proxy for the urgency in a conversation. To understand how Suicide watch subreddit users responds to a *OP* compared to other sub-reddit threads on the frontpage, we calculate differences between the posting times between consecutive messages in a reply graph. Figure 4.1a shows comparison using CDFs of inter-message response times for SW and FP threads. It can be seen that SW *OP* are responded with the highest urgency amongst the 4, especially compared to either the *OP* or any other users or sub-reddits.

How symmetric are the interactions?

Despite signs of urgency and engagement, we ask the question: what percentage of conversations happening on these subreddits are symmetric in nature ? For this The median value for U_{sym} for SW is 20% where as for AS is 0%. This shows that SW subreddit engages is a lot more symmetric conversation that the baseline threads. If we define a set of users who engage in symmetric activity with the *OP* , it would be worth while to investigate how much of the total message activity on the thread is carried out by these set of symmetric users . To calculate this we find the fraction of messages on each thread written as part of this symmetric conversation. Figure 4.1b shows the trend. It can be see that SW threads contain a higher prevalence of symmetric message exchanges compared to the baseline

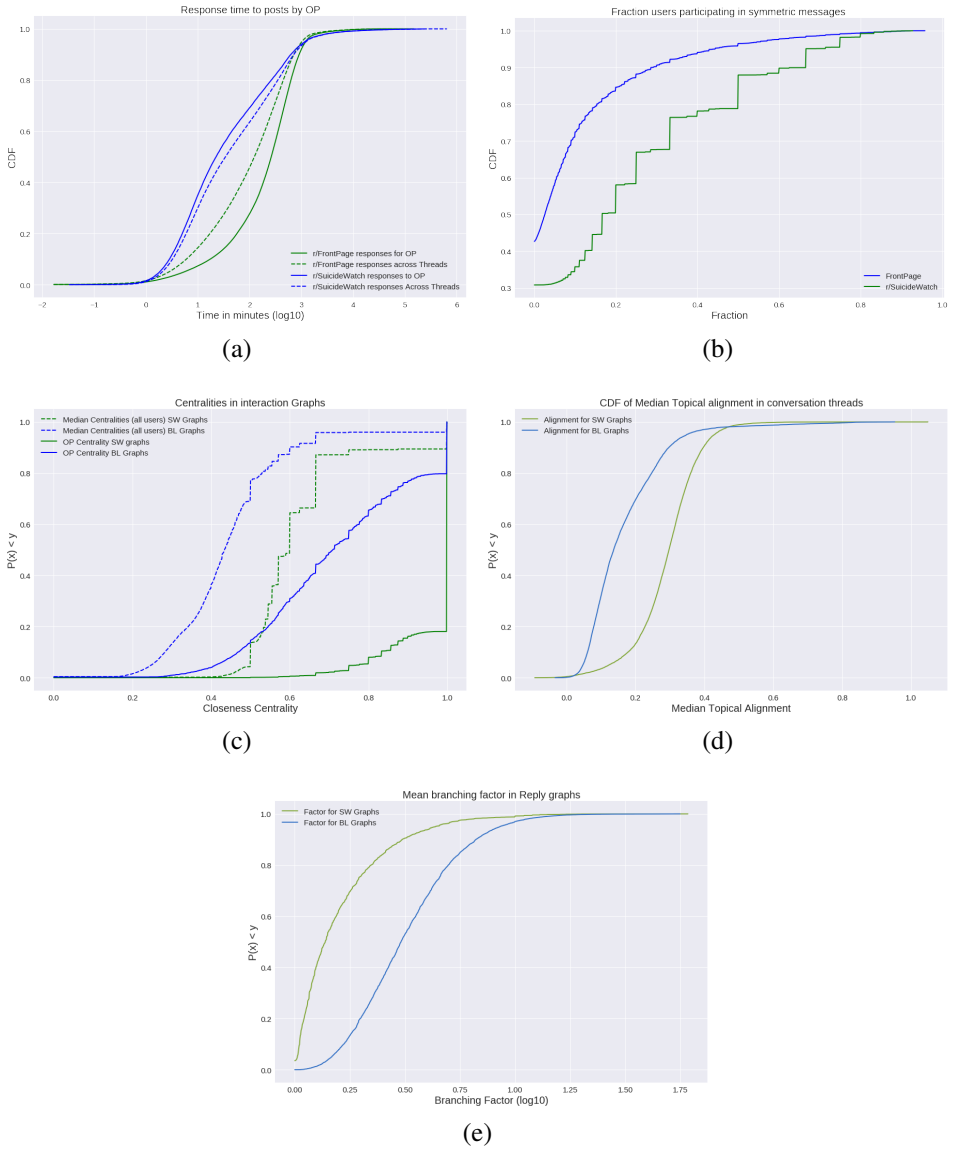


Fig. 4.1 Panel shows CDFs of different network metrics. Fig.4.1a shows the response time distributions, Fig.4.1b shows symmetrically engaged users, Fig.4.1d shows topical similarities across posts and 4.1e shows the branching factors of reply graphs.

Frontpage threads. This shows a higher engagement from the *OP*s side when participating in a supportive conversations

How central are the users?

To understand how embedded is the *OP* in a conversation thread, we compare the betweenness centralities of *OPs* in the *SW* dataset with the baseline *FP* dataset. Betweenness centrality

is a good proxy of understanding how closely linked is a node with the rest of the network. When we calculate this metric for the user graphs we see that Suicide watch *OP*s tend to have highest centralities compared to generic *FP* threads moth in terms of *OP* centrality as well as median centrality across all the users. The high centrality of *OP*s in *SW* conversations implies a high level of embedded-ness as well as a *OP* centric approach by other participants in the conversation. The Figure 4.1c shows the Empirical CDFs of centralities.

How semantically aligned are the responses?

We measure semantic alignment based on word embeddings of the source post and the reply post, at every edge of the reply graph. The detailed method of extracting semantic alignment along a post and its response is described in Section 4.3.3. Extracting such similarity metrics, we compare the trend in response text being in semantic alignment with the parent text in the reply graphs.

How branched do conversations become?

Branching in a conversation thread could be either a sign of digression or a sign interestingness resulting in more people joining in. To measure this phenomena, use the reply graphs, which resemble a n-ary directed acyclic graph, to evaluate the branching factor. By using the method described in Section 4.3.3, we found that Suicide watch threads, tend to branch in a much less as compared to our baseline conversations. This implies that suicide watch threads tend to remain a one-on-one conversation with the *OP* albeit many such dialogues may emerge, and hence that explains the high centrality of the *OP* in all interaction graphs. If the helpers on a thread seldom interact with each other, the corresponding interaction graph will have the *OP* as the most central node.

4.2.2 Patterns in local interactions

It is often a useful tool to express large interaction graphs, as the sum of local interactions between two or three nodes at a time. Such analysis is quite useful in expressing local structures in the graphs and has been used in several network analysis works For this reason we use a method more commonly known as network motif analysis (described in Section 4.3.4) to understand triads, or groups of three nodes, and the patterns of edges that exist between them.

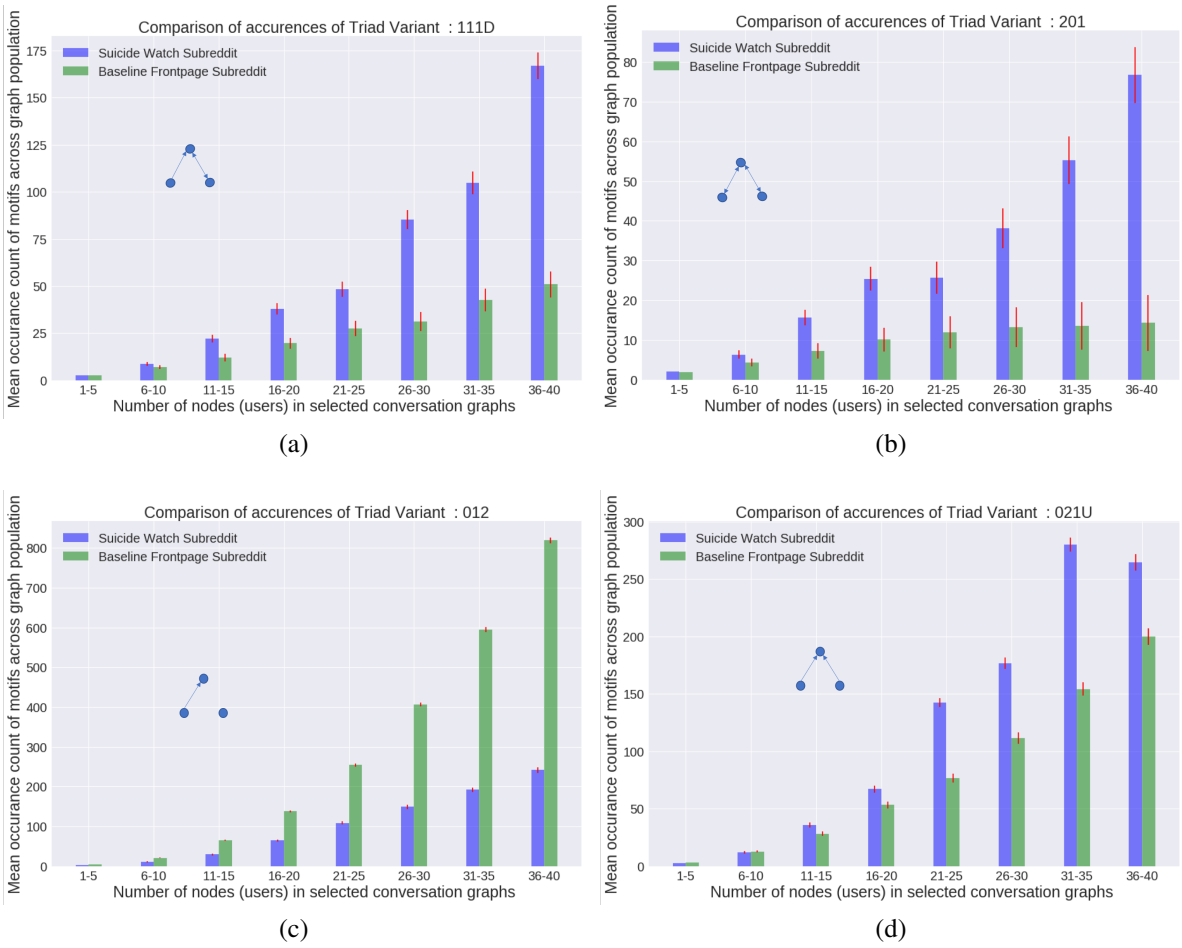


Fig. 4.2 This panel shows the statistical significance of the three over expressed and one under expressed triadic motif.

4.3 Methods

This section discusses the methodological devices used to extract insights from the fora data.

4.3.1 Data

We build on dataset that was used in [21] where they analyze textual content for the root posts in a Subreddit called Suicide Watch[5]. The dataset contains a dump of 53 thousand posts from the suicide watch sub-reddit. However the dataset did not contain the threaded conversations for each thread. Reddit is a platform where a user can create a post on a sub-reddit, to which several members of a given sub-reddit can interact with. The array of interactions may range

[5]https://www.reddit.com/r/SuicideWatch/

from simple up or down votes or posting at different hierarchy of the thread. This creates a hierarchical threaded structure of posts where the conversations are organized as threads of posts. To understand the deeper structure that is present in these posts, we crawl Reddit to get the threaded conversations by pursuing each conversation at arbitrary depth. [6]. This results in a dataset of over 50 thousand threads totaling to around 500,000 individual posts on those threads.

To baseline our work and compare theorized supportive nature of conversations with the broader community, we also crawl other reddit threads. To avoid any bias towards a particular type of subreddit, which have their own culture, we acquire roughly 50 thousand baseline posts which have been popular enough to land on the front page [7]. We crawl the Frontpage posts for 2 weeks accumulating over 50 thousand reddit threads in the process. The median amount of responses for a Suicide watch thread were 6 and for baseline Frontpage posts were 8. To understand the structure of these two foras, and find discriminating factors between a supportive community like suicide watch and a general thread on Reddit, we need to build abstractions of the thread.

4.3.2 Abstractions

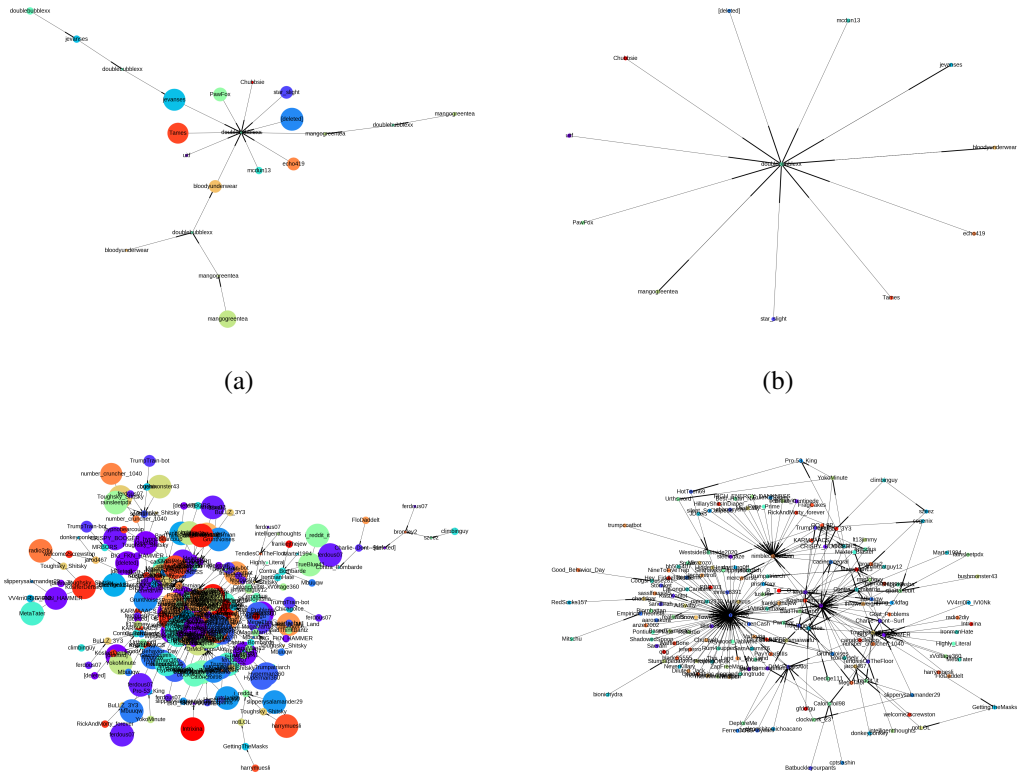
To understand the dynamics of supportive conversations, we first need to formalize the abstraction of networked conversations. In case of forum based platforms where users interact in a nested dialogue fashion, and original poster or *OP* posts a start of a thread. This thread is then open for comments by all the community users. In case of Reddit, such a community is called a Subreddit, which is a moderated collection of users who subscribe to it. These users may post new threads onto the subreddit as far as the post follows the subreddit rules. Enforcement of these rules is the responsibility of the moderators. The user who starts a thread is called the Original Poster or **OP** and the headlining post which the *OP* begins with is called the Root Post or *RP*.

Reply Graphs

The first abstraction mimics directly the structure of conversation threads on Reddit. These abstractions are called Reply Graphs. We formulate a reply graph $R\{P,E,W\}$ as a thread of multi-layered posts in a thread in response to the root post *RP* in the sub-reddit. Each graph R consists of posts $P_i, P_j, i, j \in N$, where N+1 is the total number of responses in the thread and Edges E_{ij}such that and Edge E_{ij} exists iff post P_i was in response to post P_j in

[6]The code to crawl reddit for threads can be found at *https://github.com/sagarjoglekar/redditTools*

[7]The reddit front page algorithm is a combination of popularity and decay in popularity as a function of time. More can be found here https://goo.gl/uVdHjn



(a) (b)

(c) (d)

Fig. 4.3 Example UserGraphs and their corresponding Reply graphs, Figure 4.3b shows a random thread from the SW sub-reddit and 4.3a shows the corresponding reply graph that arises from the response structure of the same thread. In comparison we have Usergraph Fig 4.3d and its corresponding reply graph Fig 4.3c from one of the Front page threads

the hierarchy of responses. The weight of the edge E_{ij} is found by calculating the cosine similarity between topic vector T_i for post P_i and the topic vector T_j of post P_j. For a given dataset, the topic vectors are extracted using the model trained on that particular corpus (LDA_{BL}, LDA_{SW}). This abstraction works well in modeling the conversational nature of these forums. For convenience of the reader, we present a couple of example pairs from SW and Frontpage baseline datasets in Figure 4.3

Interaction Graphs

In this method, we represent each thread as a directed graph $G\{V,E,W\}$ where V is the set of all users participating in a particular thread and E are the directed edges which correspond to interactions between two users $V_i, V_j \in V$. The weight of each directed edge E_{ij} corresponds

to the average of all the edges between $V_i, V_j \in V$ in the corresponding reply graph $R\{P,E,W\}$ as described above. This means that each reply graph is then mapped to a User graphs where the nodes are Users rather than posts. Another salient distinction between the two abstractions is that reply graphs resemble an n-ary tree and user graphs are directed cyclic graphs.

4.3.3 Macro and local metrics

The abstractions are used to extract certain structural metrics from the conversation threads. These metrics are then used to validate structural differences between supportive conversations and generic casual conversations from out baseline set.

Centrality

For this metric we use the User Graphs. Node centrality is a metric that measures how central a node is in a network. It directly reflects the importance of the node when it comes to membership of the shortest connecting paths between all the nodes in the graph. More formally, we use betweenness centrality of a node which is defined as Betweenness centrality of a node v is defined as

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}(v)$ is the total number of shortest paths from node s to node t and σ_{st} is the number of those paths that pass through v. To understand whether the thread starters (OP) have a special place in the network, we evaluate both OP centrality as well as median centrality across all the nodes in a User graph.

Symmetrical users

We define a symmetric user and a symmetric edges for user graphs. For a user V_i in the user graph $G\{V,E,W\}$ as described in Section , a symmetric user is a user who interacts with V_o or the OP and receives a response back from the OP. We find the fraction

$$U_{sym} = \frac{total\ number\ of\ symmetric\ users}{Total\ users\ in\ a\ thread}$$

Urgency

To understand how Suicide watch subreddit users responds to the OP and each other, compared to other sub-reddit threads on the frontpage, we calculate differences between the

posting times between consecutive messages in a reply graph. We compute the median response times per thread, for *OP*'s posts and for all other posts.

Branching Factor

Branching factor is a quantity that reflects the fan out of a conversation as it evolves. To measure this phenomena, we use the reply graphs, which resemble a n-ary directed acyclic graph, to evaluate the branching factor. The branching factor is formally described as

$$\tau = \frac{1}{|D|} \sum_{d \in D} \frac{1}{|N_d|} \sum_{n \in N_d} InDeg(n)$$

Response distance

A user who interacts frequently with a thread, may contribute at different stages of hierarchy. Intuitively, for a one on one conversation to exist, the user needs to contribute back to the thread at alternate levels of hierarchy. To measure this phenomenon, we assume that a user U contributes in a reply graph at variable depths $d_i \forall i \in [0, D]$. We calculate the average difference between two consecutive contributions d_{avg}

$$d_{avg} = \frac{\sum_{i=0}^{D} d_{i+1} - d_i}{D}$$

We calculate the values of d_{avg} for both *OP* and other users in the thread.

Semantic similarity

We use a popular word embedding method called *Word2Vec* [46] which learns representations of a set of words from a corpus of text, which in our case is the text from Suicide Watch and baseline fora. These representations can be used to extract text embedding vectors for each post which belong to a N dimensional space R^N. These vectors are tested for their alignment using cosine distance in R^N, which corresponds to semantic similarity in the textual space. This method is quite popular and used in community based question answering[45], Medical semantic similarity [16] and other medical informatics applications[86].

4.3.4 Structural metrics

Network motifs are local sub-networks between 2 or 3 nodes. Such local patterns are highly useful in quantifying local interactions and the resulting macro structure of the network[47]. They gave been used in a variety of applications and networks, from economics [83] to

cellular protein-protein interaction networks [80]. Hence we calculate census of 16 distinct *triadic motif* i.e. possible subgraphs that could be formed between any three chosen nodes.

This probably needs to be briefly introduced/explained earlier in the manuscript? To understand relation of local structures in interaction networks within a conversation with its , we compare the quantity called *motif occurrence ratio*. After calculating motif census across the dataset[4], we select progressive subsets of graphs from both datasets with nodes $n \in [k, k+4] \; \forall \; k \in \{1, 6, 11, 16, 21, 26, 31, 36\}$. This segmentation of the dataset is called binning, and it allows us to not only measure the differences in the census between the datasets, but also observe the trend as the size of the interaction graph becomes larger. We stop at 40 because beyond that, we start getting lower and lower number of graphs samples per bin.

We test each graph for the frequency of occurrence of the 16 possible triadic motifs as shown in Figure4.4a. We start by selecting the subset of graphs from Suicide watch and Frontpage belonging to a particular bin k. We then define the motif occurrence ratio as the fraction values for $\frac{\gamma_{BL}}{K_{BL}}$ and $\frac{\gamma_{SW}}{K_{SW}}$ for all the 16 motifs across all the chosen values of n.

4.4 Appendix

4.4.1 Triadic statistics for twitter conversations

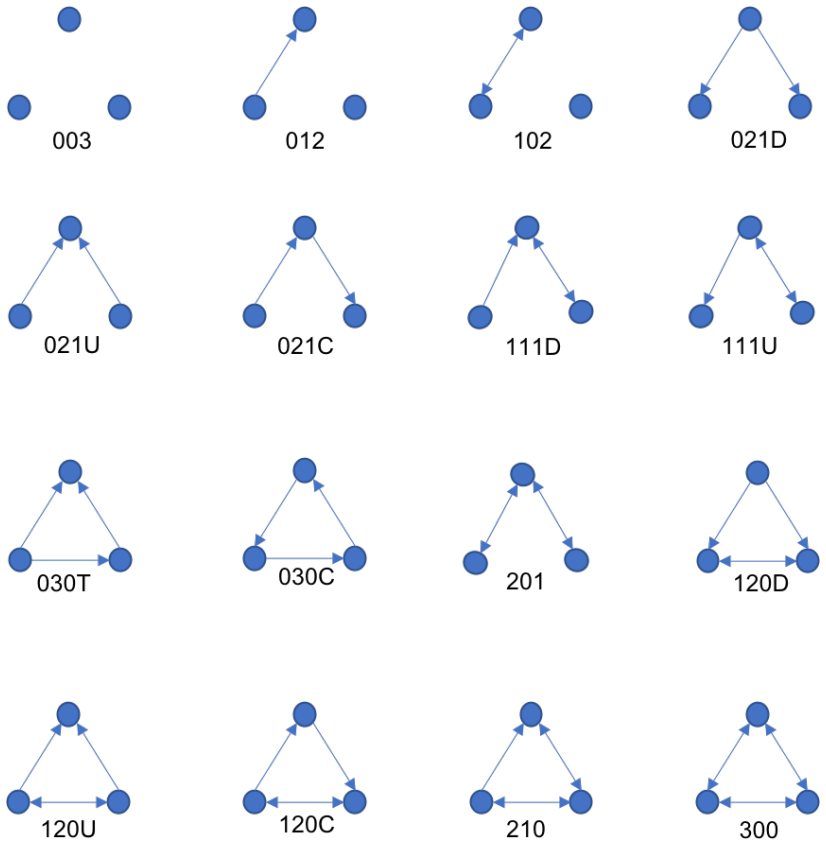
4.4.2

Replication of quantification of the topological metrics for twitter conversations for suicidal ideation and comparison with baseline threads. There are 5k threads for suicidal ideation and 6k for baseline. The baseline threads deal with discussion around westminster and manchester terrorist attacks in the uk.

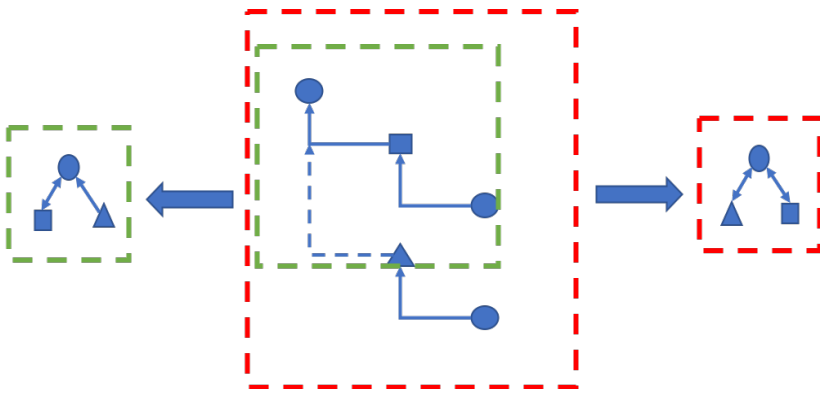
4.4.3 Network characteristics

Figure 4.7a shows the distribution of maximum depths across all Reply graphs for SW and Baseline subreddits. The SW threads depths have a median depth of 2 and mean of 4 compared to median depth of 2 for BL and a mean of 2.5. This shows that statistically the depths of Suicide watch and baseline graphs are quite similar.

Figure4.8 shows the CDF for the number of responses a Root post gets on a thread across the whole dataset. Figure 4.8 shows the CDF for number of comments per thread across the r/SuicideWatch subreddit dataset and the crawled frontpage subreddit.



(a)



(b)

Fig. 4.4 Figure 4.4a shows the 16 different types of motifs that are looked for in the user graph data. Figure 4.4b shows how three unique users could produce different motifs. The three shapes represent different users and the dotted line means the message order is irrelevant.

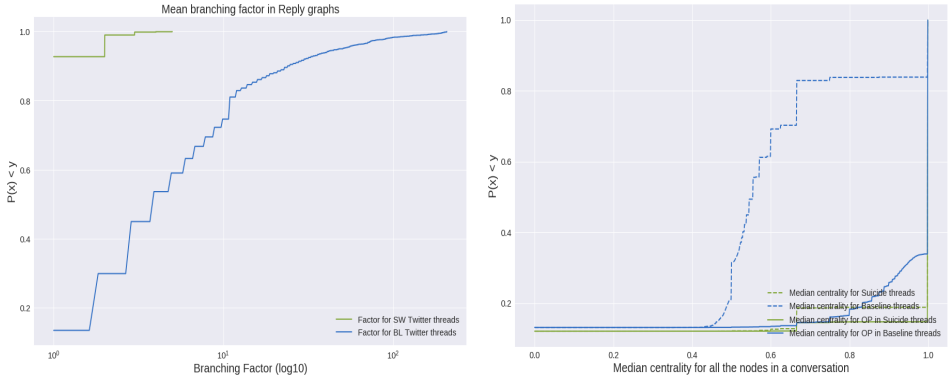
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

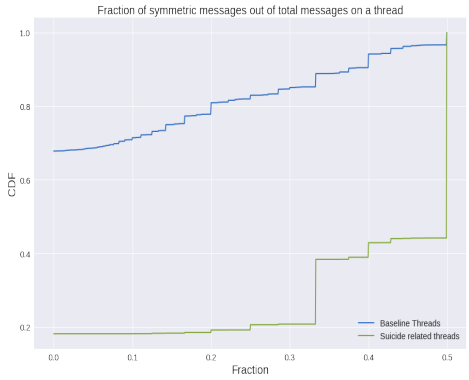
(j) (k) (l)

Fig. 4.5 The figure shows comparison of occurrence ratios of 9 insignificant motifs. Blue traces are for Suicide watch and Green traces are for Baseline Front page threads

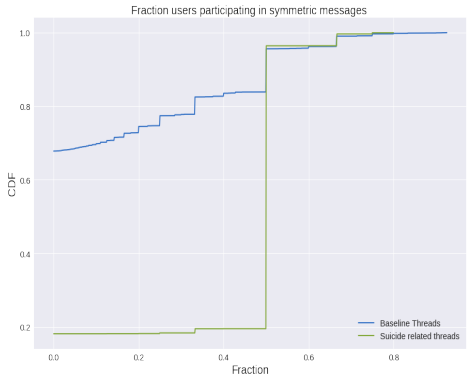


(a)

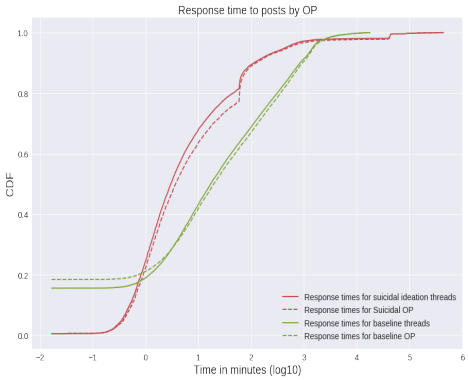
(b)



(c)



(d)



(e)

Fig. 4.6 *Fig 4.6a shows the branching factor for twitter threads that talk about suicidal tendencies against baseline threads. Fig 4.6b shows the distribution of median centralities per thread, for both the twitter crawls. Fig 4.6c shows Distribution of symmetric messages in reply graphs for both datasets. Fig 4.6d shows the distributions for users participating in a symmetric conversation Fig 4.6e shows he dustribution of reply urgency for suicide threads against baseline. The suicide median reponse time for suicide threads is 3 min as compared to 18 mins for non-suicide threads*

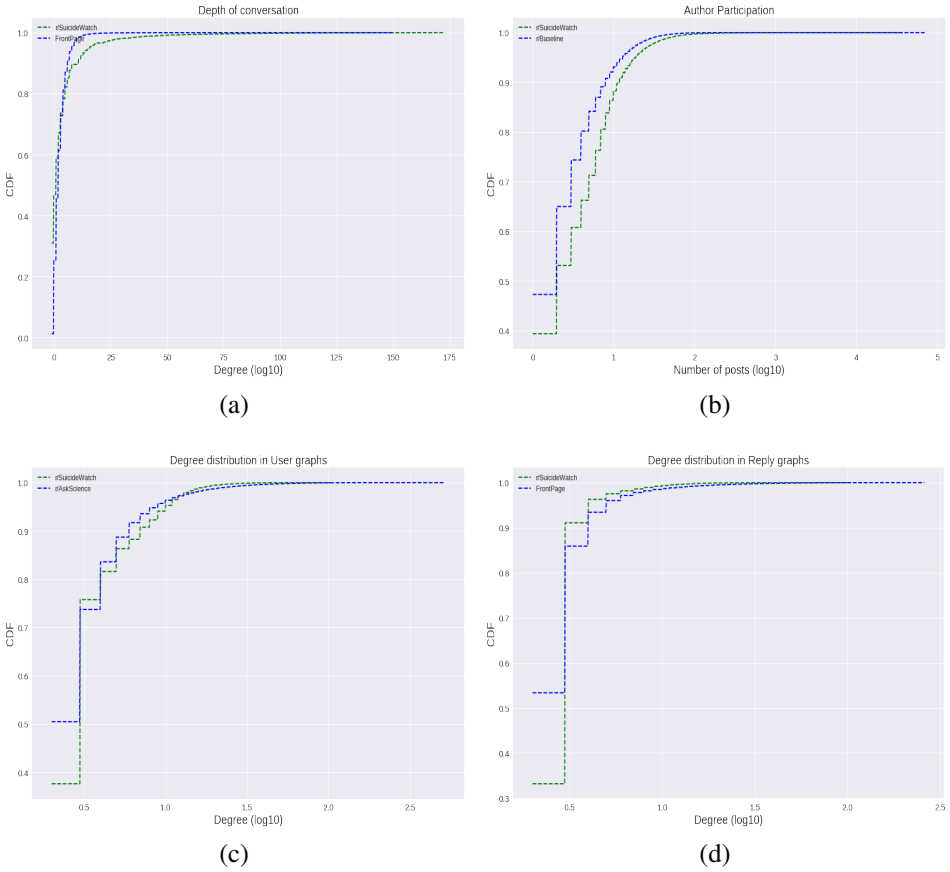


Fig. 4.7 *Fig 4.7a shows the distribution of maximum depths of Reply Graphs for Subreddit r/SuicideWatch and the baseline Frontpage conversations. Fig 4.7b shows the distribution of unique authors per thread in the two datasets. Fig 4.7d shows Distribution of degrees for Reply Graphs, r/SuicideWatch and FrontPage. Fig 4.7c shows the degree distributions for the reply graphs*

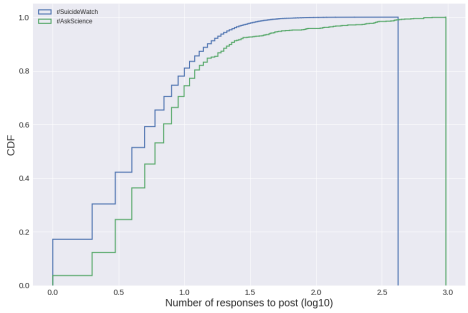


Fig. 4.8 *Distribution of responses per thread on Subreddits r/SuicideWatch and Frontpage*

CHAPTER 5

PERCEPTIONS IN REAL SPACES

There has been an explosive growth of deep learning technologies and their competency in the recent years, resulting in cross disciplinary use cases of deep learning enabled tools. In the area of computer vision and urban informatics, deep learning techniques have recently been used to predict whether urban scenes are likely to be considered beautiful, and it turns out that these techniques do so quite accurately. However, the technology falls short when it comes to generating actionable insights for AI assisted urban design. To support urban interventions, one needs to go beyond *predicting* beauty, and tackle the challenge of *recreating* beauty and *explaining* the predictors of beauty. Unfortunately, deep learning techniques have not been designed with that challenge in mind. Given their "black-box nature", they cannot even explain why a scene has been predicted to be beautiful. To partly fix that, we propose a deep learning framework (which we name FaceLift) that is able to both *beautify* existing Google Street views and *explain* which urban elements make those transformed scenes beautiful. To quantitatively evaluate our framework, we cannot resort to any existing metric (as the research problem at hand has never been faced before) and need to formulate new ones. These new metrics should ideally capture the presence (or absence) of elements that make urban spaces great. Upon a review of the urban planning literature, we identify four main metrics: walkability, green, openness, and visual complexity. For all the four metrics, the beautified scenes meet the expectations set by the literature on what great spaces tend to be made of. The transformations and their explanations are also found to be very helpful in understanding interventions for beautification, which we validate using a 20-participant expert survey. These results suggest that, in the future, as our framework's components are further researched and become better and more sophisticated, it is not hard to imagine technologies that will be able to accurately and efficiently support architects and planners in the design of the spaces we intuitively love.

5.1 Introduction

Whether a street is considered beautiful is subjective, yet research has shown that there are specific urban elements that are universally considered beautiful: from greenery, to small streets, to memorable spaces [1, 59, 62]. These elements are those that contribute to the creation of what the urban sociologist Jane Jacobs called 'urban vitality' [28].

Given that, it comes as no surprise that computer vision techniques can automatically analyse pictures of urban scenes and accurately determine the extent to which these scenes are, *on average*, considered beautiful. Deep learning has greatly contributed to increase these techniques' accuracy [17].

However, urban planners and architects are interested in urban interventions and, as such, they wish to go beyond technologies that are only able to predict beauty scores. These interests stem from the fact that the spaces we live in can be linked with several aspects of human life such as mental health[67], inequality [62] or cultural shifts [25]. They often called for technologies that would make easier to recreate beauty in urban design [12]. Deep learning, by itself, is not fit for purpose. It is not meant to recreate beautiful scenes, not least because it cannot provide any explanation on why a scene is beautiful.

To partly fix that, we propose a deep learning framework (which we name FaceLift) that is able to both *generate* a beautiful scene (or, better, *beautify* an existing scene) and *explain* why that scene is beautiful. This opens up a possibility of using technology in urban planning efforts like decision making based of subjective opinions, participatory urban planning and promotion of restorative urban design such as green spaces and walkable areas. Through this work, we make two main contributions:

- We propose a deep learning framework that is able to learn whether Google Street views are beautiful or not, and that, based on that training, is able to both *beautify* existing views and *explain* which urban elements make these views beautiful (Section 5.3).

- We quantitatively evaluate whether the framework is able to actually produce beautified scenes (Section 5.4). We do so by proposing a family of four urban design metrics that we have formulated based on a thorough review of the literature in urban planning. For all these four metrics, the framework passes with flying colors: with minimal interventions, beautified scenes are twice as walkable as the original scenes, for example. Also, after building an interactive tool with "FaceLifted" scenes in Boston and presenting it to twenty experts in architecture, we found that the majority of them agreed on three main areas of our work's impact: decision making, participatory urbanism, and promotion of restorative spaces among the general public.

For sake of brevity, we will use the term 'Urban Scene' through out the paper to address an arbitrary Google Street View image. The image is fetched from a particular latitude and longitude point on the map. In the rest of the paper we explore related literature across various tracks of urban perceptions and urban beauty in Section 5.2. We then describe in detail the Facelift framework in Section 5.3. The evaluation of the framework is described in detail in Section 5.4 We conclude by pointing out some limitations and biases that might well guide future work (Section 5.5).

5.2 Related Work

Previous work has focused on collecting ground truth data about how people perceive urban spaces, on predicting urban qualities from visual data, and on generating synthetic images that enhance a given quality (e.g., beauty).

Ground truth of urban perceptions. So far the most detailed studies of perceptions of urban environments and their visual appearance have relied on personal interviews and observation of city streets: for example, some researchers relied on annotations of video recordings by experts [63], while others have used participant ratings of simulated (rather than existing) street scenes [40]. The web has recently been used to survey a large number of individuals. Place Pulse is a website that asks a series of binary perception questions (such as 'Which place looks safer [between the two]?') across a large number of geo-tagged images [62]. In a similar way, Quercia *et al.* collected pairwise judgments about the extent to which urban scenes are considered quiet, beautiful and happy [59]. They were then able to analyze the scenes together with their ratings using image-processing tools, and found that the amount of greenery in any given scene was associated with all three attributes and that cars and fortress-like buildings were associated with sadness. Taken all together, their results pointed in the same direction: urban elements that hinder social interactions were undesirable, while elements that increase interactions were the ones that should be integrated by urban planners to retrofit cities for greater happiness.

Deep learning and the city. Computer vision techniques have increasingly become more sophisticated. Deep learning techniques, in particular, have been recently used to accurately predict urban beauty [17, 68], urban change [48], and even crime [15]. These works also did some interesting analysis of the data to understand how safety, depression, beauty and other such dimensions are perceived across urban spaces. [17] also utilized deep learning methods to train models capable of comparing two urban images for their perception values in terms

Symbol	Meaning
I_i	Original urban scene
Y	Set of annotation classes for urban scenes (e.g., beautiful, ugly)
y_i	Annotation class in Y (e.g., beautiful)
$\hat{I}_j$	Template scene (synthetic image)
I'	Target Image
C	Beauty Classifier

Table 5.1 Notations

of beauty et.al. However even these works did not dive into the reasoning aspect of these models.

Generative models. Deep learning has recently been used not only to analyze existing images but also to generate new ones. n the past couple of years, there have been papers which exploit generative version of neural nets to delve into the aspects of explainability. Nguyen *et al.* [49] used generative networks to create a natural-looking image that maximizes a specific neuron. In theory, the resulting image is the one that "best activates" the neuron under consideration (e.g., that associated with urban beauty). In practice, it is still a synthetic template that needs further processing to look realistic.

To sum up, a lot of work has gone into collecting ground truth data about how people tend to perceive urban spaces, and into building accurate predictions models of urban qualities. However, little work has gone into models that generate realistic urban scenes and that offer human-interpretable explanations of what they generate.

5.3 FaceLift Framework

The goal of FaceLift is to take as input a geo-located urban scene and give as output its transformed (beautified) version. To that end, it performs in five steps:

- **Curating urban scenes** It is common knowledge that deep learning systems need immense amount of data. In this first step we try to develop a sound framework for curating and augmenting annotated images, on which the model could be trained.

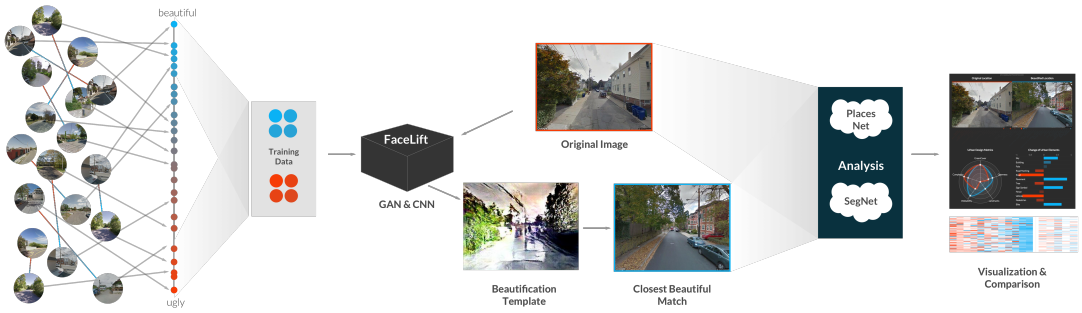


Fig. 5.1 A simplistic end to end illustration of the FaceLift framework.

- **Training a beauty classifier** To generate beauty, you first need a reliable model that could learn the representation of beauty. To achieve this, we train a deep learning model that could distinguish beautiful urban scenes from non-beautiful urban scenes.

- **Generating a synthetic beautified scene** Based on the learned representation of beauty, we train a Generative model which could augment the beauty of an input urban scene.

- **Retrieving a realistic beautified scene** as showcased in Figure 5.1, the generated images are representations of beautified input urban scene in a latent space. This latent representation needs to be transformed back to a realistic looking image, using retrieval.

- **Identifying the urban elements characterizing the beautified scene** In the final step, the framework explains changes introduced in the transformation process in terms of literature-driven urban design metrics, and quantifies these changes as metrics for urban beauty.

Step 1 Curating Urban Scenes

To begin with, we need highly curated training data with labels reflecting urban beauty. We start with the Place Pulse dataset that contains 100k Google Street Views across 56 cities around the world [17]. These scenes are labeled in terms of whether the corresponding places are likely to be perceived beautiful, depressing, rich, and safe. We focus only on those scenes that are labeled in terms of beauty and that have at least three judgments. This leave us with roughly 20,000 scenes. To transform judgments into beauty scores, we use the TrueSkill algorithm [24], which gives us a way of partitioning the scenes into two sets (Figure 5.2): one containing beautiful scenes, and the other containing ugly scenes. The resulting set of scenes

Augmentation	Accuracy (Percentage)
None	63
Rotation	68
Rotation + Translation	64
Rotation + Conservative Translation	73.5

Table 5.2 Percentage accuracy for our beauty classifier trained on differently augmented sets of urban scenes.

is too small for training any deep learning module without avoiding over-fitting though. As such, we need to augment such a set.

We do so in two ways. First, we feed each scene's location into the Google Streetview API to obtain the snapshots of the same location at different camera angles (i.e., at $\theta \in -30^\circ, -15^\circ, 15^\circ, 30^\circ$). However, the resulting dataset is still too small for robust training. Therefore, again, we feed each scene's location into the Google Streetview API, but now we do so to obtain other scenes at distance $d \in \{10, 20, 40, 60\}$ meters. This will greatly expand our set of scenes, but it might do so at the price of introducing scenes whose beauty scores have little to do with the original scene's. To fix that, we take only the scenes that are *similar* to the original one (we call this way of augmenting "conservative translation"). To compute the similarity between a pair of scenes, we represent the two scenes with visual features derived from the FC7 layer of PlacesNet and compute the similarity between the two corresponding feature vectors [85]. For all scenes at increasing distance $d \in \{10, 20, 40, 60\}$ meters, we take only those whose similarity scores with the original scene is above a threshold. In a conservative fashion, we choose that threshold to be the median similarity between rotated and original scenes (those of the first augmentation step).

To make sure this additional augmentation has not introduced any unwanted noise, we consider two sets of scenes: one containing those that have been taken during this last step, i.e. the one with high similarity to the original scenes (*taken-set*), and the other containing those that have been filtered away (*filtered-set*). Each scene is then scored with PlacesNet [85] and is represented with the five most confident scene labels. We then aggregate labels at set level, by computing each label's frequency on the *taken-set* and on the *filtered-set*. Finally, we characterize each label's propensity to be correctly augmented as: $\text{prone}(label) = fr(label, taken\text{-}set) - fr(label, filtered\text{-}set)$. This reflects the extent to which a scene with a given label is prone to be augmented or not. From Figure 5.4, we find that, as one would expect, scenes that contain highways, fields and bridges can be augmented at increasing distances while still showing resemblances to the original scene; by contrast, scenes that contain gardens, residential neighborhoods , plazas, and skyscrapers cannot be easily augmented, as they are often found in high density parts of the city, where there is tremendous diversity within short distances.

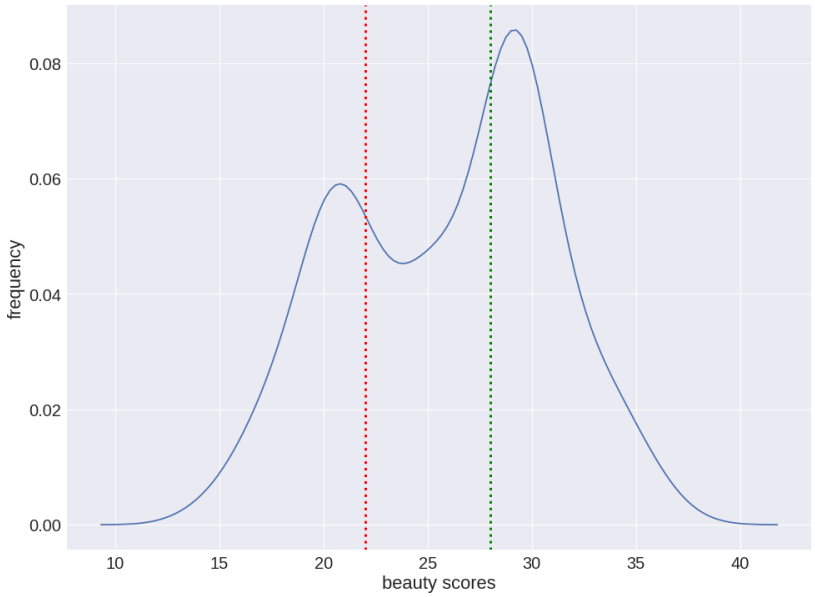


Fig. 5.2 Frequency distribution of beauty scores. The red and green lines represent the thresholds below and above which images are considered ugly and beautiful. Conservatively, images in between are discarded.

(a) (b)

Fig. 5.3 Two types of augmentation: (a) rotation of the Street Views camera (based on rotation); and (b) exploration of scenes at increasing distances (based on translation).

Step 2 Training a beauty classifier

Having this highly curated set of labeled urban scenes, we are now ready to train classifier C with labels reflecting our beauty assessments. We use the CaffeNet architecture, a modified

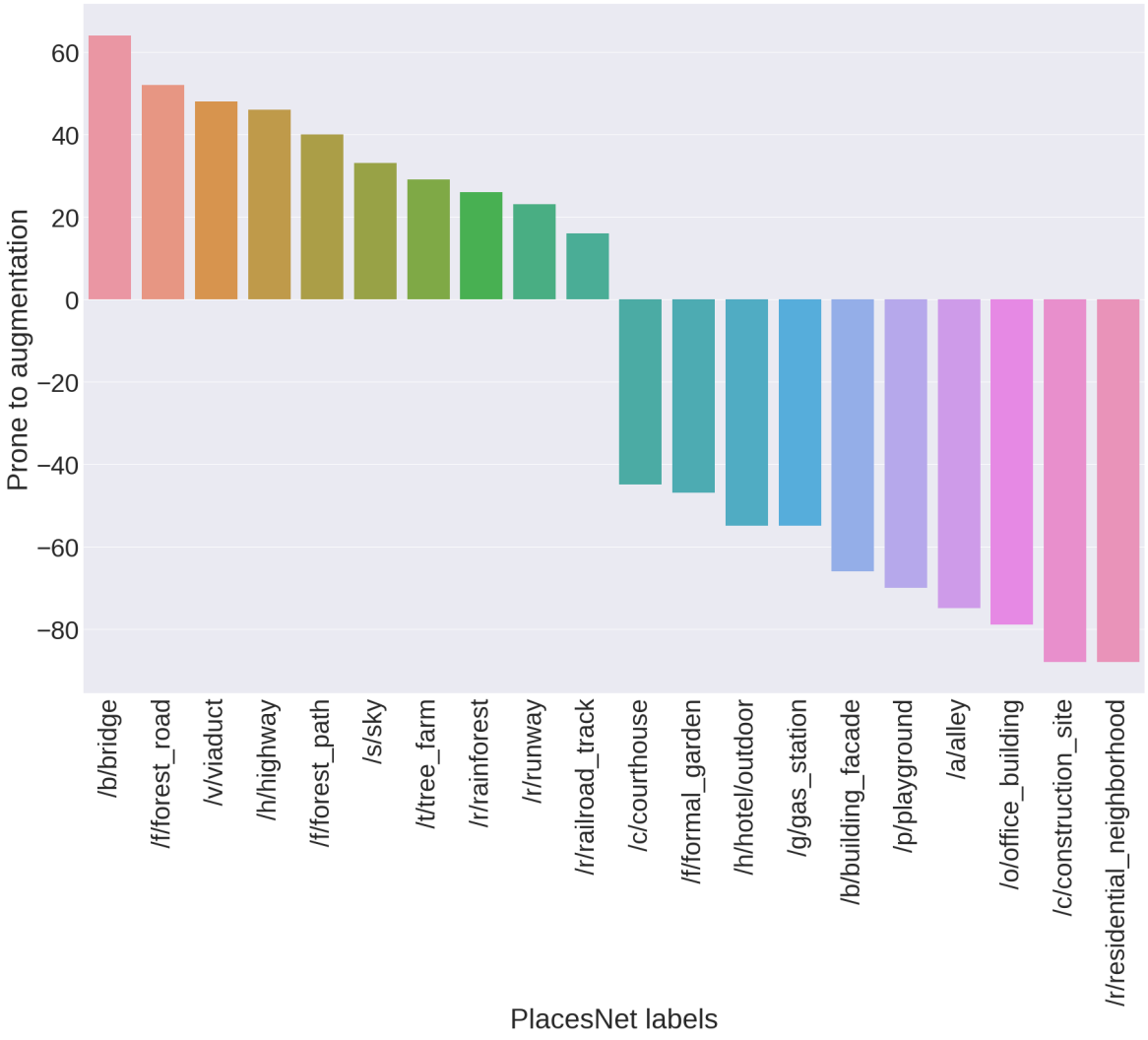


Fig. 5.4 The types of scene that have greater propensity to be correctly augmented with similar scenes at increasing distances.

version of AlexNet [35, 74]. The training is done on a 70% split of the data, and the testing on the remaining 30%. All this is done on increasingly augmented sets of data. We start from our 20k images and progressively augment them with the snapshots obtained with the 5-angle camera rotations, and then with the exploration of scenes at increasing distance $d \in \{10, 20, 40, 60\}$ meters. The idea behind data augmentation is that accuracy would increase with it. Indeed it does (Table 5.2): it goes from 63% on the set of original scenes to as much as 73.5% on the set of fully augmented scenes, which is a notable increase in accuracy for such classes of classification tasks. As a baseline, we compare with the models trained by Dubey et.al in [17] on the same seed data that we use for our pipeline. They report that their models perform at 70% accuracy in the task of picking a beautiful image amongst

any two given images. Albeit the set-up of our model is not to compare two images but just to classify a particular image in a binary class, this baseline shows that our model is showing a comparable performance in beauty classification.

Step 3 Generating a synthetic beautified scene

Having this trained classifier at hand, we can then build a generator of synthetic beautified scenes. This is a model that, given the two classes ugly y_i and beautiful y_j, transforms any original scene I_i of class y_i (e.g., ugly scene) into template scene $\hat{I}_j$ that maximizes class y_j (e.g., beautified template scene).

More specifically, given an input image I_i known to be of class y_i (e.g., ugly), our technique outputs $\hat{I}_j$, which is a more beautiful version of it (e.g., I_i is morphed towards the average representation of a beautiful scene) while preserving I_i's details. The technique does so using the "Deep Generator Network for Activation Maximization" (*DGN-AM*) [49]. Given an input image I_i, *DGN-AM* iteratively re-calculates the color of I_i's pixels in a way the output image $\hat{I}_j$ both maximizes the activation of neuron y_j (e.g., the "beauty neuron") and looks "photo realistic", which is done by conditioning the maximization to an "image prior". This is equivalent to finding the feature vector f that maximizes the following expression:

$$\hat{I}_j = G(f) : \arg\max_f (C_j(G(f)) - \lambda ||f||) \tag{5.1}$$

where:

- $G(f)$ is the image synthetically generated from the candidate feature vector f;
- $C_j(G(f))$ is the activation value of neuron y_j in the scene classifier C (the value to be maximized);
- λ is a L_2 regularization term.

Here the initialization of f is key. If f were to be initialized with random noise, then the resulting $G(f)$ would be the average representation of category y_j (of, e.g., beauty). Instead, since f is initialized with I_i, then the resulting $G(f)$ is I_i's version "morphed to become more beautiful".

Step 4 Returning a realistic beautified scene

We now have template scene $\hat{I}_j$ (which is a synthetic beautified version of original scene I_i) and need to retrieve a realistic looking version of it. We do so by: *i)* representing each of our

original scenes in Step 1 (including $\hat{I}_j$) as a 4096 dimensional feature vector derived from the FC7 layer of the PlacesNet [85]; *ii)* computing the distance (as L_2 Norm) between $\hat{I}_j$'s feature vector and each of the original scene's feature vector; and *iii)* selecting the original scene most similar (smaller distance) to $\hat{I}_j$. This results into the selection of the beautified scene I_j.

Step 5 Identifying characterizing urban elements

Since original scene I_i and beautified scene I_j are real scenes and we make sure that they maintain the same structural characteristics (e.g., point of view, layout), we can easily compare them in terms of presence or absence of SegNet's and PlacesNet's labels. That is, we can determine how the original scene and its beautified version differ in terms of urban design elements. This step required us to develop metrics inspired from urban design literature, to quantify the changes in elements. A detailed description of the characterization and evaluation would follow in Section 5.4

5.4 Evaluation

The goal of FaceLift is to transform existing urban scenes into versions that: *i)* people perceive more beautiful; *ii)* contain urban elements typical of great urban spaces; *iii)* are easy to interpret; and *iv)* architects and urban planners find useful. To ascertain whether FaceLift meets that composite goal, we answer the following questions next:

Q1 Do individuals perceive"FaceLifted" scenes to be beautiful?

Q2 Does our framework produce scenes that possess urban elements typical of great spaces?

Q3 Which urban elements are mostly associated with beautiful scenes?

Q4 Do architects and urban planners find FaceLift useful?

Q1 People's perceptions of beautified scenes

To ascertain whether FaceLifted scenes are perceived by individuals as they are supposed to, we run a crowd-sourcing experiment on Amazon Mechanical Turk. We randomly select 200 scenes, 100 beautiful and 100 ugly (taken at the bottom 10 and top 10 percentiles of the Trueskill's score distribution of Figure 5.2). Our framework then transforms each ugly scene into its beautified version, and each beautiful scene into its corresponding 'uglified'. These scenes are arranged into pairs, each of which contains the original scene and its beautified or

Original (I_i)	Latent Beauty representation ($\hat{I}_j$)	Beautified (I_j)
		
		
		
		
		

Table 5.3 The table showcases examples of the"FaceLifting" process. It is worth observing

uglified version. On Mechanical Turk, we only select verified masters for our crowd-sourcing workers (those with an approval rate above 90% during the past 30 days), pay them $0.1 per task, and ask each of them to choose the beautiful scene for given pairs. We make sure to have at least 3 votes for each scene pair. Overall, our workers end up selecting the scenes that are actually beautiful 77.5% of the times, suggesting that FaceLifted scenes are correctly perceived most of the times.

Q2 Are beautified scenes great urban spaces?

To answer that question, we need to understand what makes a space great. After a careful review of the urban planning literature, we identify four factors [1, 19] (summarized in Table 5.4): great places mainly tend to be walkable, offer greenery, feel cozy, and be visually rich.

Metric	Description
Walkability	Walkable streets increase the social capital of a place, and they appeal to the exploring nature of the human psyche [19, 58, 72].
Green Spaces	The presence of greenery has repeatedly been found to impact people's well being [1]. Under certain conditions, it could also promote social interactions [59]. This suggest that not all greenery has to be considered in the same way though: dense forests or unkempt greens might well have a negative impact [28].
Privacy-Openness	A sense of privacy (as opposed to a sense of openness) impacts a place's perception [19].
Visual Complexity	Visual complexity is a measure of how diverse a urban scene is in terms of design materials, textures, and objects [19].

Table 5.4 Urban Design Metrics

To automatically extract visual cues related to these four factors, we select 500 ugly scenes and 500 beautiful ones at random, transform them into their opposite aesthetic qualities (i.e., ugly ones are beautified, and beautiful ones are 'uglified'), and compare which urban elements related to the four factors distinguish uglified scenes from beautified ones.

We extract labels from each of our 1,000 scenes using two image classifiers. First, using PlacesNet [85], we label each of our scenes according to a classification containing 205 labels (reflecting, for example, landmarks, natural elements), and retain the five labels with highest confidence scores for the scene. Second, using Segnet [2], we label each of our scenes

according to a classification containing 12 labels. Segnet is trained on dash-cam images, and classifies each scene pixel with one of these twelve labels: road, sky, trees, buildings, poles, signage, pedestrians, vehicles, bicycles, pavement, fences, and road markings.

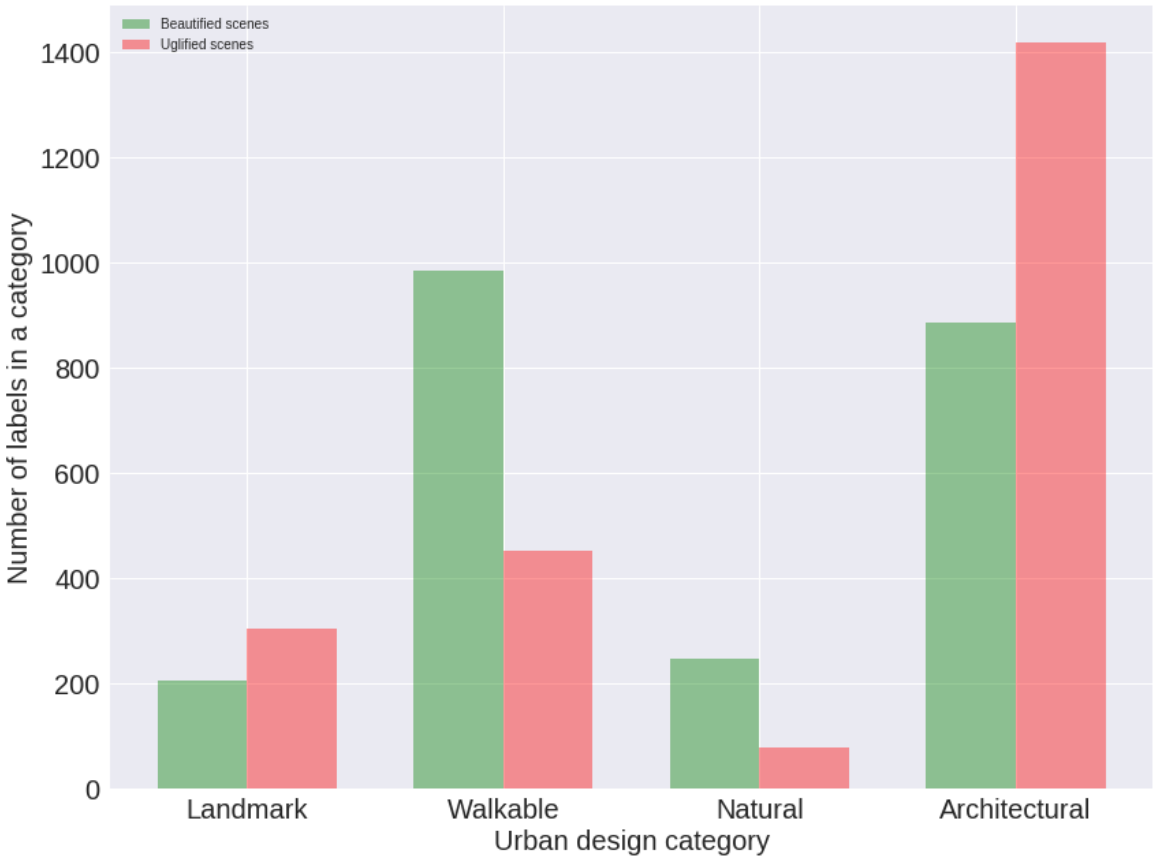


Fig. 5.5 Number of labels in specific urban design categories (on the *x*-axis) found in beautified scenes as opposed to those found in uglified scenes.

Having these two ways of labeling scenes, we can now test whether the expectations set by the literature describing metrics of great urban spaces (Table 5.4) are met in the FaceLifted scenes.

H1 Beautified scenes tend to be walkable. We manually select only the PlacesNet labels that are related to walkability. These labels include, for example, *abbey, plaza, courtyard, garden, picnic area, and park*. To test hypothesis *H1*, we count the number of walkability-related labels found in beautified scenes as opposed to those found in uglified scenes (Figure 5.5): the former contain twice as many walkability labels than the latter. We then determine which types of scenes are associated with beauty (Figure 5.6). Unsurprisingly, beautified scenes

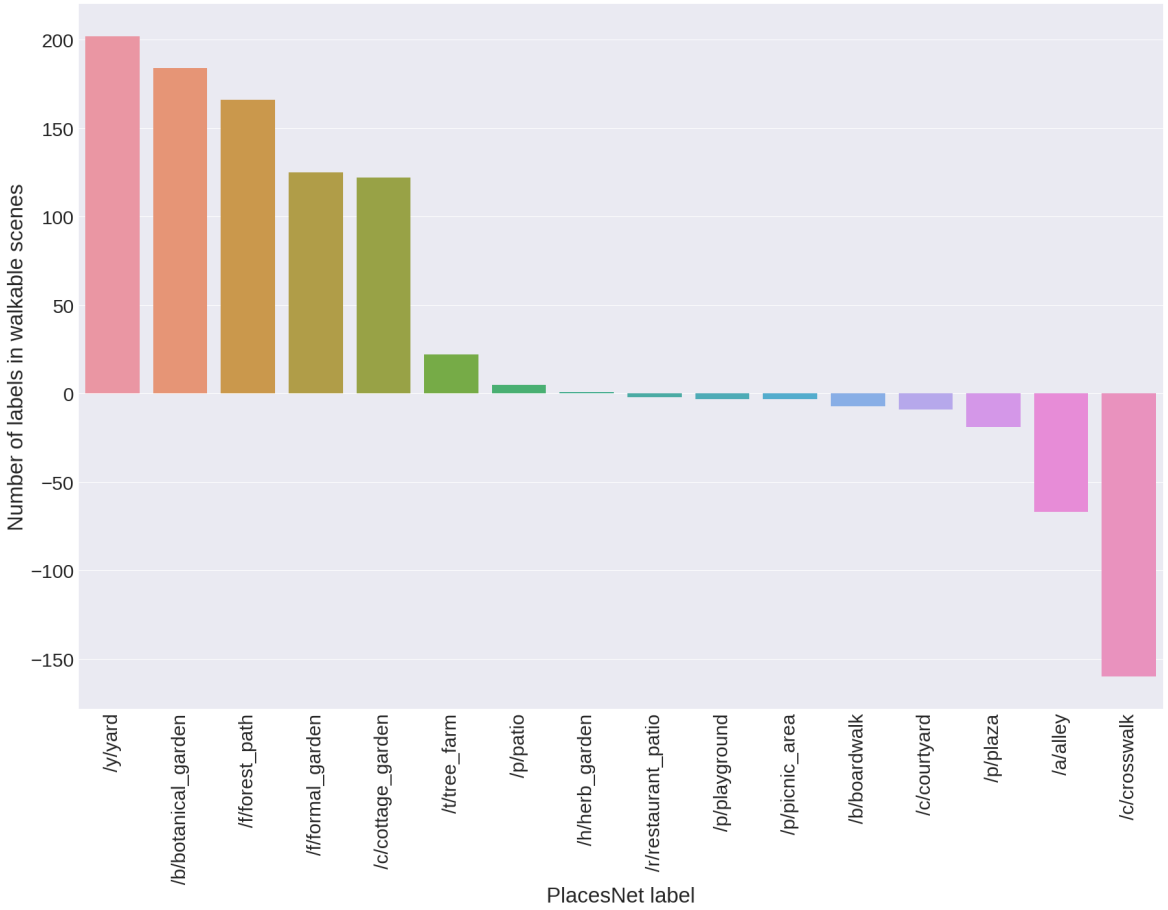
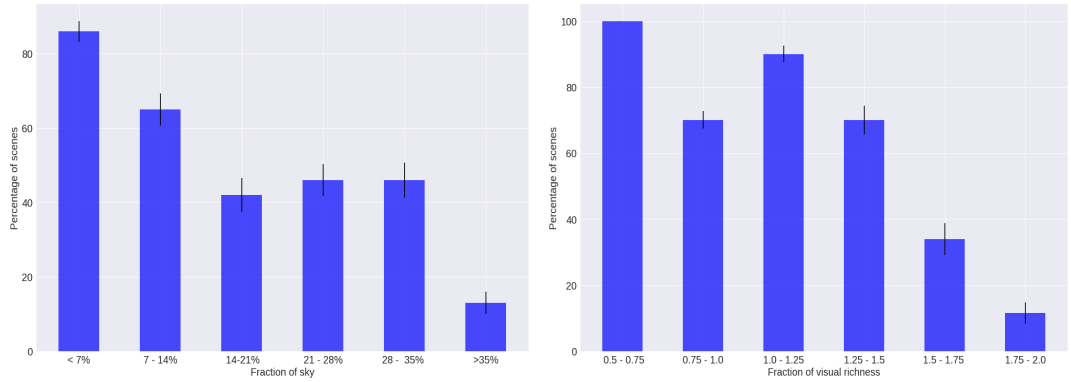


Fig. 5.6 Count of specific walkability-related labels (on the x-axis) found in beautified scenes minus the count of the same labels found in uglified scenes.

tend to show gardens, yards, and small paths. By contrast, uglified ones tend to show built environment features such as shop fronts and broad roads.

H2 Beautified scenes tend to offer green spaces. We manually select only the PlacesNet labels that are related to greenery. These labels include, for example, *fields, pasture, forest, ocean, and beach*. Then, in our 1,000 scenes, to test hypothesis *H2*, we count the number of nature-related labels found in beautified scenes as opposed to those found in uglified scenes (Figure 5.5): the former contain more than twice as many nature-related labels than the latter. To test this hypothesis further, we compute the fraction of 'tree' pixels (using SegNet's label 'tree') in beautified and uglified scenes, and find that beautification adds 32% of tree pixels, while uglification removes 17% of them.

H3 Beautified scenes tend to feel private and 'cozy'. To test hypothesis *H3*, we count the



(a) (b)

Fig. 5.7 The percentage of scenes (y-axis): (a) having an increasing presence of sky (on the x-axis); and (b) having an increasing level of visual richness (on the x-axis). The error bars represent standard errors obtained by random re-sampling of the data for 500 iterations.

fraction of pixels that Segnet labeled as 'sky' and show the results in a bin plot in Figure 5.7a: the x-axis has six bins (each of which represents a given range of sky fraction), and the y-axis shows the percentage of beautified *vs.* uglified scenes that fall into each bin. Beautified scenes tend to be cozier (lower sky presence) than the corresponding original scenes.

H4 Beautified scenes tend to be visually rich. To quantify to which extent scenes are visually rich, we measure their visual complexity [19] as the amount of disorder in terms of distribution of (Segnet) urban elements in the scene:

$$H(X) = -\sum p(i) \log p(i) \tag{5.2}$$

where i is the i^{th} Segnet's label. The total number of labels is twelve. The higher $H(X)$, the higher the scene's entropy, that is, the higher the scene's complexity. To test hypothesis *H4*, we show the percentage of scenes that fall into a complexity bin (Figure 5.7b): beautified scenes are of low to medium complexity, while uglified ones are of high complexity.

Q3 Urban elements of beautified scenes

To determine which urban elements are the best predictors of urban beauty and the extent to which they are so, we run a logistic regression, and, to ease interpretation, we do so on one pair of predictors at the time:

$$Pr(\text{beautiful}) = logit^{-1}(\alpha + \beta_1 * V_1 + \beta_2 * V_2 + \beta_3 * V_1.V_2) \tag{5.3}$$

Pair of urban elements	β_1	β_2	β_3	Error Rate (Percentage)
Buildings - Trees	-0.032	0.084	0.005	12.7
Sky - Buildings	-0.08	-0.11	0.064	14.4
Roads - Vehicles	-0.015	-0.05	0.023	40.6
Sky - Trees	0.03	0.11	-0.012	12.8
Roads - Trees	0.04	0.10	-0.031	13.5
Roads - Buildings	-0.05	-0.097	0.04	20.2

Table 5.5 Coefficients of logistic regressions run on one pair of predictors at the time.

where $V1$ is the fraction of the scene's pixels marked with one Segnet's label, say, "buildings" (over the total number of pixels), and $V2$ is the fraction of the scene's pixels marked with another label, say, "trees". The result consists of three beta coefficients: β_1 reflects $V1$'s contribution in predicting beauty, β_2 reflects $V2$'s contribution, and β_3 is the interaction effect, that is, it reflects the contribution of the dependency of $V1$ and $V2$ in predicting beauty. We run logistic regressions on the five factors that have been found to be most predictive of urban beauty [1, 19, 59], and show the results in Table 5.5.

Since we are using logistic regressions, the quantitative interpretation of the beta coefficients is eased by the "divide by 4 rule" [76]: we can take β coefficients and "divide them by 4 to get an upper bound of the predictive difference corresponding to a unit difference" in beauty [76]. For example, take the results in the first row of Table 5.5. In the model $Pr(beautiful) = logit^{-1}(\alpha - 0.032 \cdot buildings + 0.084 \cdot trees + 0.005 \cdot buildings \cdot trees)$, we can divide - 0.032/4 to get -0.008: a difference of 1 in the fraction of pixels being buildings corresponds to no more than a 0.8% *negative* difference in the probability of the scene being beautiful. In a similar way, a difference of 1 in the fraction of pixels being trees corresponds to no more than a 0.021% *positive* difference in the probability of the scene being beautiful. By considering the remaining results in Table 5.5, we find that, across all pairwise comparisons, trees is the most positive element associated with beauty, while roads and buildings are the most negative ones. Since these results go in the direction one would expect, one might conclude that the scenes beautified by our framework are in line with previous literature, adding further external validity to our work.

Q4 Do architects and urban planners find it useful?

To ascertain whether practitioners find FaceLift potentially useful, we built an interactive map of the city of Boston in which, for selected points, we showed pairs of urban scenes before/after beautification (Figure 5.8). We then sent that map along with a survey to 20 experts in architecture, urban planning, and data visualization around the world. The experts had to complete tasks in which they rated FaceLift based on how well it supports decision

Use case	Definitely Not	Probably Not	Probably	Very Probably	Definitely
Decision Making	4.8%	9.5%	38%	28.6%	19%
Participatory Urban Planning	0%	4.8%	52.4%	23.8%	19%
Promote Green Cities	4.8%	0%	47.6%	19%	28.6%

Table 5.6 Urban experts polled about the extent to which an interactive map of "FaceLifted" scenes promotes: (a) decision making; (b) citizen participation in urban planning; and (c) promotion of green cities

Fig. 5.8 Interactive map of FaceLifted scenes in Boston.

making, participatory urbanism, and promotion of green spaces among the general public. The results are show in Table 5.6 according to our experts, the tool can very probably supports

decision making, probably support participatory urbanism, and definitely promote green spaces. These results are qualitatively supported by our experts' comments, which include: "*The maps reveal patterns that might not otherwise be apparent*", "*The tool helps focusing on parameters to identify beauty in the city while exploring it*", and "*The metrics are nice. It made me think more about beautiful places needing a combination of criteria, rather than a high score on one or two dimensions. It made me realize that these criteria are probably spatially correlated*".

5.5 Conclusion

FaceLift is a transparent framework that beautifies existing urban scenes. This translates into two main technical advancements. First, FaceLift is able to generate realistic scenes as opposed to existing approaches based on Generative Adversarial Networks whose final transformations are quite coarse as they still take the form of synthetic templates. Second, it augments the deep learning black-box with a module that offers explanations on what has been transformed, making that box more transparent.

There are still important limitations though. One is data bias. The framework is as good as its training data, and more work has to go into collecting reliable ground truth of human perceptions. This data should ideally be stratified according to the people's characteristics that impact their perceptions. The other main limitation is that generative models are hard to control, and more work has to go into offering principled ways of fine-tuning the generative process.

Despite these limitations, FaceLift has the potential to support urban interventions in scalable and replicable ways: it can be applied to the scale of an entire city, and that can be replicated in other cities. The advantage of shifting the focus of research away from predictive analytics towards urban interventions is that people could be part of discussions on works of architecture more than they are nowadays. To turn existing spaces into something more beautiful, that will still be the duty of architecture. Yet, with technologies similar to FaceLift more readily integrated in the architecture discussions, the complex job of recreating restorative spaces in an increasingly urbanized world will be greatly simplified. After all, "we delight in complexity to which genius have lent an appearance of simplicity." [12] In the context of future work, that genius is represented by the future technologies that we will contribute to build to deal with the complexity of our cities.

CHAPTER 6

PERCEPTION AND GENERATIVE MODELS

REFERENCES

[1] Alexander, C., Ishikawa, S., Silverstein, M., i Ramió, J. R., Jacobson, M., and Fiksdahl-King, I. (1977). *A pattern language*. Gustavo Gili.

[2] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*.

[3] Bakhshi, S., Shamma, D. A., and Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 965–974. ACM.

[4] Batagelj, V. and Mrvar, A. (2001). A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks*, 23(3):237–243.

[5] Borges, M. (2004). What can kant teach us about emotions? *The Journal of Philosophy*, 101(3):140–158.

[6] Cecchi, L., Liccardi, G., Pellegrino, F., and Sofia, M. (2012). 2 social networks: A new source of psychological stress or a way to enhance self-esteem? negative and positive implications in bronchial asthma. *Journal of Investigational Allergology and Clinical Immunology*, 22(6):402.

[7] Cha, M. et al. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th WWW*, WWW '09, pages 721–730, New York, NY, USA. ACM.

[8] Chen, J. et al. (2016). Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 898–907, New York, NY, USA. ACM.

[9] Crawford, M. (2015). *The World Beyond Your Head: How to Flourish in an Age of Distraction*. penguin publishing.

[10] Datta, R. et al. (2008). Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE ICIP*, pages 105–108. IEEE.

[11] Davenport, T. H. and Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business Press.

[12] De Botton, A. (2008). *The Architecture of Happiness*. Vintage Series. Knopf Doubleday Publishing Group.

[13] De Choudhury, M. and De, S. (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media*, pages 71–80.

[14] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA. ACM.

[15] De Nadai, M., Vieriu, R. L., Zen, G., Dragicevic, S., Naik, N., Caraviello, M., Hidalgo, C. A., Sebe, N., and Lepri, B. (2016). Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *Proceedings of the ACM on Multimedia Conference (MM)*.

[16] De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., and Bruza, P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822. ACM.

[17] Dubey, A., Naik, N., Parikh, D., Raskar, R., and Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. *arXiv preprint arXiv:1608.01769*.

[18] et.al, K. (2015). How to take a good selfie? In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 923–926, New York, NY, USA. ACM.

[19] Ewing, R. and Clemente, O. (2013). *Measuring urban design: Metrics for livable places*. Island Press.

[20] Fontanini, G. et al. (2016). Web video popularity prediction using sentiment and content visual features. In *Proceedings of the 2016 ACM on ICMR*, pages 289–292. ACM.

[21] Gkotsis, G. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7:45141.

[22] Grossman, D. (2013). Can micro video change how we communicate? BBC Newsnight.

[23] Hare, J. S. and others. (2011). Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM MM*, MM '11, pages 691–694, New York, NY, USA. ACM.

[24] Herbrich, R., Minka, T., and Graepel, T. (2007). Trueskill™: a bayesian skill rating system. In *Advances in neural information processing systems*, pages 569–576.

[25] Hristova, D., Aiello, L. M., and Quercia, D. (2018). The new urban success: How culture pays. *Frontiers in Physics*, 6:27.

[26] Hwang, K. O., Ottenbacher, A. J., Green, A. P., Cannon-Diehl, M. R., Richardson, O., Bernstam, E. V., and Thomas, E. J. (2010). Social support in an internet weight loss community. *International journal of medical informatics*, 79(1):5–13.

[27] Isola, P., Xiao, J., Torralba, A., and Oliva, A. (2011). What makes an image memorable? In *CVPR, 2011 IEEE Conference on*, pages 145–152. IEEE.

[28] Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House.

[29] Joglekar, S., Sastry, N., and Redi, M. (2017). Like at first sight: Understanding user engagement with the world of microvideos. In *International Conference on Social Informatics*, pages 237–256. Springer.

[30] Jou, B. et al. (2015). Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM MM*, pages 159–168. ACM.

[31] Kant, I. (1987). *Critique of judgment*. Hackett Publishing.

[32] Kavuluru, R., Ramos-Morales, M., Holaday, T., Williams, A. G., Haye, L., and Cerel, J. (2016). Classification of helpful comments on online suicide watch forums. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '16, pages 32–40, New York, NY, USA. ACM.

[33] Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. In *2006 IEEE CVPR'06*, volume 1, pages 419–426. IEEE.

[34] Khosla, A. et al. (2014). What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, New York, NY, USA.

[35] Krizhevsky, A. et al. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105.

[36] Kummervold, P. E., Gammon, D., Bergvik, S., Johnsen, J.-A. K., Hasvold, T., and Rosenvinge, J. H. (2002). Social support in a wired world: use of online mental health forums in norway. *Nordic journal of psychiatry*, 56(1):59–65.

[37] Lartillot, O. and Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244.

[38] Laurier, C., Lartillot, O., Eerola, T., and Toiviainen, P. (2009). Exploring relationships between audio features and emotion in music.

[39] Leung, L. (2009). User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. *New media & society*, 11(8):1327–1347.

[40] Lindal, P. J. and Hartig, T. (2012). Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*.

[41] Louppe, G. et al. (2013). Understanding variable importances in forests of randomized trees. In *NIPS*, pages 431–439.

[42] Luo, Y. and Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399. Springer.

[43] Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM MM*, MM '10, New York, NY, USA. ACM.

[44] Mazloom et al. (2016). Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, New York, NY, USA.

[45] Mihaylov, T. and Nakov, P. (2016). Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 879–886.

[46] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[47] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

[48] Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., and Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576.

[49] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016a). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395.

[50] Nguyen, P. X., Rogez, G., Fowlkes, C., and Ramamnan, D. (2016b). The open world of micro-videos. *arXiv preprint arXiv:1603.09439*.

[51] Nwana, A. O. et al. (2013). A latent social approach to youtube popularity prediction. In *2013 IEEE (GLOBECOM)*, pages 3138–3144. IEEE.

[52] O'Brien, H. L. and Toms, E. G. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

[53] Park, A. and Conway, M. (2018). Harnessing reddit to understand the written-communication challenges experienced by individuals with mental health disorders: Analysis of texts from mental health communities. *J Med Internet Res*, 20(4):e121.

[54] Park, A., Conway, M., and Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Computers in human behavior*, 78:98–112.

[55] Perlovsky, L. (2014). Aesthetic emotions, what are their cognitive functions? *Frontiers in Psychology*, 5:98.

[56] Pinto, H. et al. (2013). Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM ICWSM*, pages 365–374. ACM.

[57] Pogue, D. (2013). Why are micro movies so popular these days? *Scientific American*.

[58] Quercia, D., Aiello, L. M., Schifanella, R., and Davies, A. (2015). The Digital Life of Walkable Streets. In *Proceedings of the 24th ACM Conference on World Wide Web (WWW)*, pages 875–884.

[59] Quercia, D., O'Hare, N. K., and Cramer, H. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 945–955. ACM.

[60] Redi, M. and Others (2014). 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE CVPR*, pages 4272–4279.

[61] Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, 33(2):163–180.

[62] Salesses, P., Schechtner, K., and Hidalgo, C. A. (2013). The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400.

[63] Sampson, R. J. and Raudenbush, S. W. (2004). Seeing Disorder: Neighborhood Stigma and the Social Construction of Broken Windows. *Social Psychology Quarterly*, 67(4).

[64] Schifanella, R. et al. (2015). An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of THE 9TH ICWSM 2015*.

[65] Schifanella, R. et al. (2016). Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM MM*, pages 1136–1145. ACM.

[66] Schupp, H. T., Flaisch, T., Stockburger, J., and Junghöfer, M. (2006). Emotion and attention: event-related brain potential studies. *Progress in brain research*, 156:31–51.

[67] Seresinhe, C. I., Preis, T., and Moat, H. S. (2015). Quantifying the impact of scenic environments on health. *Scientific reports*, 5:16899.

[68] Seresinhe, C. I., Preis, T., and Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7):170170.

[69] Slater, A. and Kirby, R. (1998). Innate and learned perceptual abilities in the newborn infant. *Experimental Brain Research*, 123(1-2):90–94.

[70] Soat, M. (2015). Social media triggers a dopamine high. *Marketing News*, 49(11):20–21.

[71] Song, Y., Dixon, S., and Pearce, M. (2012). A survey of music recommendation systems and future perspectives. In *9th International Symposium on Computer Music Modeling and Retrieval*, volume 4.

[72] Speck, J. (2012). Walkable City: How Downtown Can Save America, One Step at a Time. In *Farrar, Straus and Giroux*.

[73] Squire, M. (2015). " should we move to stack overflow?" measuring the utility of social media for developer support. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 2, pages 219–228. IEEE.

[74] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[75] Totti et al. (2014). The impact of visual attributes on online image diffusion. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, New York, NY, USA. ACM.

[76] Vaughn, B. K. (2008). Data analysis using regression and multilevel/hierarchical models, by gelman, a., & hill, j. *Journal of Educational Measurement*, 45(1):94–97.

[77] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

[78] Wang, Y. et al. (2015). Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *Ninth ICWSM*.

[79] Yamasaki, T. et al. (2014). Social popularity score: Predicting numbers of views, comments, and favorites of social photos using only annotations. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, WISMM '14, New York, NY, USA. ACM.

[80] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences*, 101(16):5934–5939.

[81] Yeh, C.-H. e. a. (2010). Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM MM*, pages 211–220. ACM.

[82] Zadra, J. R. and Clore, G. L. (2011). Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, 2(6):676–685.

[83] Zhang, X., Shao, S., Stanley, H. E., and Havlin, S. (2014). Dynamic motifs in socio-economic networks. *EPL (Europhysics Letters)*, 108(5):58001.

[84] Zhong, C. et al. (2015). Predicting pinterest: Automating a distributed human computation. In *Proceedings of the 24th WWW*, WWW '15, New York, NY, USA. ACM.

[85] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.

[86] Zhu, Y., Yan, E., and Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC medical informatics and decision making*, 17(1):95.