METHODS AND RESULTS SECTION

MODEL

Our model builds on research across sociology, psychology, economics and philosophy on the idea of self-concept and self-presentation. This paper posits that individuals present different 'selves' in different situations (Mead's (1934), Goffman (1975)). Online platforms, such as dating sites, provide an avenue to strategically construct a different 'self' from what may be observed or perceived in real life.

The text section of these profiles provides considerable freedom in the process of this construction. It may prove critical in cases where other aspects of the profile- such as the photo and fixed-response answers may not project an attractive image as desired by the user. For the sake of this study, we assume that the variables where such a dichotomous allocation of profiles into 'attractive' and 'not attractive' exists would include height, education, weight and ethnicity. Past research (see literature review) suggests that female users accord much greater importance to this written component of men's profiles than men do to women's. Hence, we will limit ourselves to men's profiles.

Individuals who consider themselves to be 'deficient' or somehow lacking in the four listed dimensions may then rely on the use of words to reinforce their strengths or possibly even present a factually untrue version of themselves. A large volume of such essays could then serve as 'documents' for computational methods to discern major topics and themes that may be seen to emerge across the consolidated corpus from all essays.

Thus, this research seeks to identify and analyze any substantial differences across the categories of the above variables in the proportion of men using these topics. For example, our models expects a different distribution across topics for short and non-short men, or men with only high-school education versus those with more. For preliminary testing of this line of inquiry, we will only be considering one of the four variables- height.

Another key aspect of the model is that of preferences being determined by opportunity in terms of the status quo (Bruch, 2019). In the context of online dating, preferences can be seen to be constructed in terms of what is seen repeatedly. In this sense, the distribution of topics across profiles would paint an image of what is increasingly perceived as 'normal'. Whether a preference for this norm develops cannot be realistically determined with this date. Nonetheless, we will be able to test which subgroups of individuals aim to either confirm or deviate from these norms- perhaps as an extension of Heino' (2010)'s self marketing.

In a related vein, selectivity about mates has typically been observed more in women than men (Finkel et al, 2011). The setting of high expectations could act as a signal of high demand by other women, and the ability to choose. It would also reflect self-awareness. Alternatively, many responses may reflect the woman's preferences- that the man is on board provided that the woman is.

DATA

This study will leverage publicly open and anonymized user profile data for 59,946 users of OkCupid within a 25 mile radius of San Francisco that were extracted with permission (Kim & Escobedo-Land, 2012). To ensure that these were active users, profiles were accepted into dataset only if they had been members as of 06/26/2012, had been active in the previous year, and had at least one photo in their profile. The data set includes "typical user information, lifestyle variables, and text responses to 10 essay questions"

https://github.com/rudeboybert/JSE_OkCupid

Unlike several studies in this domain, this set does not contain details of interactions between users. In this sense, it limits our scope of testable hypotheses, since we cannot explore which profiles 'triggered' other users to interact with them, or sustain that interaction either online, offline or both. Nonetheless, it does provide a critical window into the realm of self-concept and self-presentation, which forms the focus of this paper.

The data were themselves submitted by users in the San Francisco Bay Area for the purposes of the dating app. In this sense, they suffer from self-reporting bias, especially in some of the areas where individuals are likely to veer towards dishonesty in their response so as to improve their dating prospects. This is an inherent limitation of the data. Data can be triangulated as a whole- say for the distribution of heights within the sample vis-a-vis census records. However, they cannot be verified for each individual. In subsequent analysis, we may make adjustments to account for actual height distribution. For now, we will trust the veracity of these claims.

After their submission, the data were then scraped but with anonymization and removal of personally identifiable information for legal and privacy concerns. We dropped all female users, and male users with missing entries for any of the 4 key variables or essays. This reduced the total sample size to 25,281.

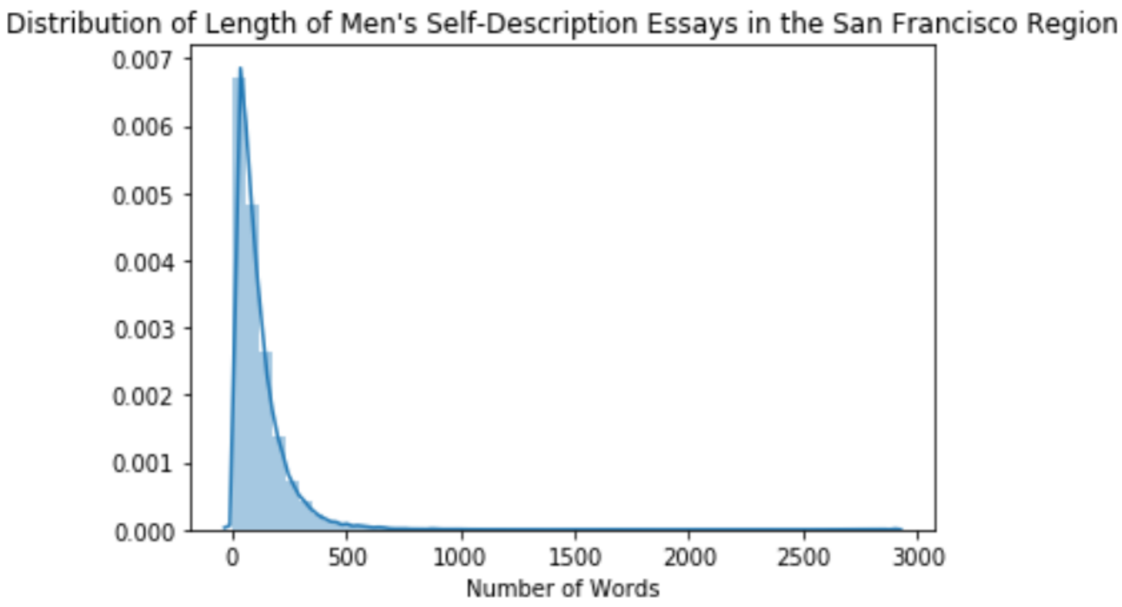
This data includes 9 optional essays available in this data. While all of them enquire into different aspects of self-concept, the first one- entitled 'About Me'- most closely matches the kind of presentation this study is interested in. In real-life situations, it is the first to be displayed, and thus builds the crucial first impression. Subsequent essays may not even be browsed if this first one fails to hold a viewer's attention. For these reasons, we will only rely on the first essay to link to the theoretical construct of self-concept. Later work will then explore aspects of selectivity through Essay 9'0 'You Should Message Me If'

The remaining variables link directly to their stated role in the our theoretical model. We will limit the number of categories under each to more of a dichotomy, as mentioned in the model introduction. This can be developed into a more fine-grained analysis at later stages.

SUMMARIZING KEY VARIABLES

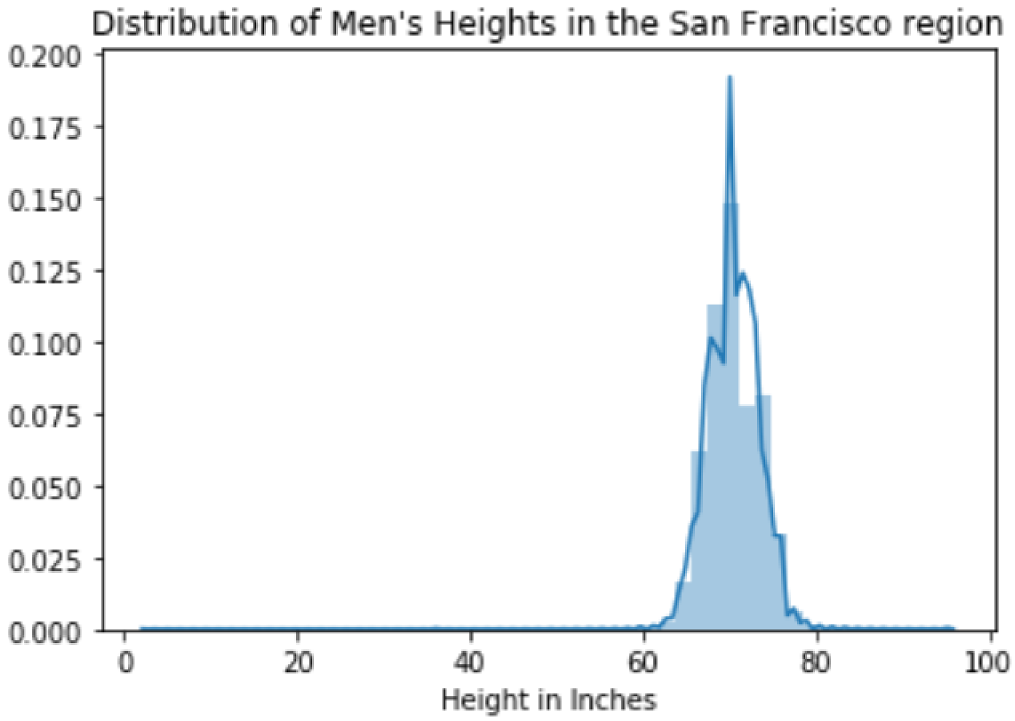
We begin with an exploration of the two continuous variables in this model- the essay length and height. The y axis represents the proportion of the total counts represented by that particle value (on the x axis) for a given variable.

Essay Length



Distribution of Length of Men's Self-Description Essays in the San Francisco Region

Height



Distribution of Men's Heights in the San Francisco region

Key Statistics for Continuous Variables

Statistic	Essay Length (in words)	Height (in inches)
mean	109.251058	70.462165
std	113.561951	3.012987
min	1	3

25%	39	69
50%	78	70
75%	143	72
max	2894	95

The short size of the average introduction essay holds interesting ramifications for this research. Fewer topics may be included within each essay. The long tail for this distribution may also open up questions on why certain users invest noticeably greater numbers of words into their self-presentation. This analysis would serve as a supplement to topic modelling.

The statistics for the height variable confirm our hunch from the visualization that height is largely normally distributed, while essay size is heavily skewed.

The quartiles distribution also helps define the cutoff for what will qualify as ‘short’ in this study: the 1st quartile. Thus, male users under 69 inches will form the ‘short’ category, while everyone else will fall in the ‘not short’ category. The use of this classification results in the following distribution of counts:

Height Type	Count	Proportion
Not Short	17467	63.90
Short	9870	36.10
Total	27337	100.00

Though constituting only about a third of the users, the ‘short’ category still represents a sufficient number to able to discern any pertinent trends in self-description. We will now explore this aspect in the next section.

METHODS

The key methods harnessed here are topic modelling applied to different populations a tfidf (Term Frequency- Inverse Document Frequency) matrix and use of NMF (Non-NegativeMatrix Factorization) to cluster the data points into 25 groups. The choice of Vocabulary includes unigrams, bigrams, and trigrams without redundancies.

The study will begin with a cleaning of the data. All HTML tags, such as break lines and hyperlinks would need to be removed. Spelling discrepancies will not be corrected for, as the level of language fluency and correctness may factor into how such profiles are evaluated.

We now have a corpus through this collection of text documents. This corpus may, in turn, be represented as a matrix of token counts which weighs each term by its absolute frequency. Extremely common terms- such as article (‘a’, the’) in English do not provide any new informations. Thus, instead of raw counts, tokens may be weighted through term frequency-inverse document frequency (tf-idf). These are well defined in Shishido (2016):

“Tf-idf is the product of the term frequency and the inverse document frequency. The term frequency refers to the relative frequency of term t in document d. The inverse document frequency is the log of the total number of documents N to the number of documents that contain term t.”

In the final analysis, we will finalize the same list of topics as in Shishido (2016), and then measure the proportion of shorter male users who rely on that topic. This will similarly be extended for men with less education, non-white ethnicity and unhealthy weight in consolidated visualization featuring the distinct subgroups simultaneously. For now, we simply run the NMF analysis separately on the subsets of men characterized as short, and those who are not short.

INITIAL RESULTS

We attempt the creation of two different numbers of topics- 25 and 10. Moving from the former to the latter helps understand if some overlapping themes could be compressed into one without loss of any important differences. Using 25 also helps maintain comparability with Shishido (2016), whose work was built on the same dataset.

As with any topic modelling exercise, the interpretation of what topics the words comprise remains open-ended and subjective. Nonetheless, we will aim to maintain uniformity in this judgment to the maximum extent possible.

WITH 25 TOPICS

Topics Referenced by Short Users

Topic Sr. No	Key Words	Column1	Possible Interpretation
Topic 0	don ve think ll say way make pretty lot self		Unclear
Topic 1	years moved sf school ago ve city lived california college		Relocation for Education
Topic 2	new trying places friends meeting try meet city york exploring		Exploration
Topic 3	just ask moved wanna trying got say hang hi average		Unclear
Topic 4	like watch play movies games stuff feel don hang think		Leisure
Topic 5	im dont lol chill pretty haha thats lets alot old		Laid back
Topic 6	love laugh music travel outdoors art cook sports food heart		Hobbies
Topic 7	life live living enjoying share fullest family believe world enjoy		Living It Up
Topic 8	san francisco born raised city living moved diego grew native		Origins and Moving
Topic 9	things try trying doing learning simple different enjoy tend little		Small pleasures
Topic 10	going easy laid pretty friendly laugh earth humor hang adventures		Laid back
Topic 11	good humor sense food conversation listener company times great pretty		Good Company
Topic 12	time spend free spending friends family spent lot busy having		Family and Friends
Topic 13	bay area raised born living native east grew lived transplant		Origins and Moving
Topic 14	people meet meeting laugh interesting make making cool world lot		Social Interaction
Topic 15	fun loving having outgoing likes times caring honest adventurous best		Adventure
Topic 16	work hard play week lot company balance day live tech		Work-Life Balance
Topic 17	guy nice earth laid kind chill likes pretty funny friends		Likeable

Topic 18	know want message ask don hit talk wanna send need	Outward Contact
Topic 19	looking meet relationship share date woman partner special forward girl	Relationship Needs
Topic 20	enjoy music movies friends playing watching sports hiking games cooking	Hobbies
Topic 21	open minded adventurous curious mind honest earth relationship funny shy	Personality Description
Topic 22	really shy don message trying awesome sure bad think writing	Reserve
Topic 23	working hard currently school college living playing graduated engineer law	Career
Topic 24	person loves type caring kind hang shy earth outgoing creative	Personality Description

Topics Referenced by Non-Short Users

Topic Sr. No	Key Words	Column1	Possible Interpretation
Topic 0	people don think way say ll make person self right		Unclear
Topic 1	area bay native lived grew moved recently living originally east		Relocation
Topic 2	life live enjoy world share living enjoying fullest experiences want		Living it Up
Topic 3	looking meet relationship woman share girl partner man hang forward		Relationship Needs
Topic 4	like stuff don feel watch cook movies play people read		Hobbies
Topic 5	new people meet meeting places trying city try york experiences		Novelty
Topic 6	love laugh travel outdoors cook music food especially city world		Hobbies
Topic 7	san francisco living city moved diego live originally native lived		Relocation
Topic 8	im dont lol thats chill person say hi lets kinda		Unclear
Topic 9	time spend free spending working lot having spent long busy		Busy
Topic 10	going easy person open pretty laid try minded honest funny		Personality Description
Topic 11	good food humor sense conversation company wine pretty laugh times		Fun
Topic 12	just ask trying moved chill hang say wanna got start		Outward Contact
Topic 13	ve lived got years lot traveled pretty seen bit places		Travel
Topic 14	moved sf years ago school year city grew went college		Relocation
Topic 15	work hard play working live lot week school company try		Busy
Topic 16	things try trying doing learning lot simple make learn new		Novelty
Topic 17	fun loving having outgoing adventurous person loves laugh smart likes		Fun
Topic 18	know want don ask message let ll getting talk tell		Outward Contact
Topic 19	enjoy music movies playing sports watching hiking games outdoors reading		Hobbies
Topic 20	guy nice laid pretty earth funny kind likes loves chill		Laidback
Topic 21	coast east west transplant grew bay originally school came summer		Relocation for Education
Topic 22	born raised california college went lived oakland school currently northern		Relocation for Education
Topic 23	friends family great hanging close make important loyal hang group		Social Interaction

Topic 24	really cool maybe pretty say oh feel bad don guess	Unclear

We largely see the same major themes of travel, outward contact, novelty, being busy, personality (with emphasis on novelty, fun and adventure, being laidback), social interaction and hobbies. At this granularity, there is no clear difference as yet in the topics per se. We now try to combine together some of the repeated ideas.

With 10 Topics

Topics Referenced By Short Users

Topic Sr. No	**Key Words**	**Possible Interpretation**
Topic 0	don know good people really time think ve say want make ll pretty person things way lot right kind little	Unclear
Topic 1	bay area years moved sf school ve lived college east grew born ago work city year coast living went california	Relocation
Topic 2	new people meet looking things meeting trying places try friends city enjoy learning exploring open explore york learn restaurants travel	Novelty
Topic 3	just ask want know moved looking message wanna guy ll say trying got hang hi really meet hey hit average	Outward Contact
Topic 4	like things music play movies games watch enjoy playing watching video sports going work food friends stuff hiking bike lot	Hobbies
Topic 5	im dont lol chill pretty haha thats lets alot know old shy nice really hard person hit kinda outgoing say	Laidback
Topic 6	love music food sports play travel outdoors laugh art friends movies playing good watching family cook cooking especially hiking watch	Hobbies
Topic 7	life enjoy live work things family hard living share enjoying believe world fullest friends great simple important working happy passionate	Social Relationships
Topic 8	san francisco city moved born raised living diego grew years recently lived year currently native live work home york chicago	Relocation
Topic 9	fun going easy guy looking good loving enjoy person laid earth friends nice likes open movies humor loves funny honest	Fun

Topics Referenced By Non-Short Users

Topic Sr. No	**Key Words**	**Possible Interpretation**

Topic 0	life don ve know people think want really way make things ll person say world self right little best good	Unclear
Topic 1	bay area years moved sf east school born lived coast grew ago raised college went ve year california originally recently	Relocation
Topic 2	fun guy looking easy going loving nice laid funny good earth life person friends likes pretty open humor sense loves	Fun & Humour
Topic 3	just ask know want moved looking really message ll trying meet say wanna hang chill maybe cool don questions right	Outward Contact
Topic 4	like things don stuff people really think going feel movies play lot watch games pretty read fun cook music cool	Hobbies
Topic 5	new people meet things meeting trying places try looking friends city learning exploring interesting experiences open york explore enjoy traveling	Novelty & Adventure
Topic 6	love music laugh travel outdoors live life food cook art sports especially nature world play great city hiking family passionate	Hobbies
Topic 7	san francisco moved city living years born ve year lived ago originally raised grew diego recently live college school currently	Relocation
Topic 8	im dont lol thats chill know person laid guy say kinda wanna hi hit lets old school hard shy really	Laidback
Topic 9	enjoy time good work friends music going playing working movies spend watching play sports food hiking games outdoors family reading	Hobbies

Again, the pattern has remained largely the same. The words themselves seem to also overlap, though this may be verified through a more detailed analysis. From this first attempt it would appear that there may exist a more useful complement to topic modelling- examining outliers. In terms of Bruch’ s (2019) model on opportunity and preferences, we may find that less common phrases and topic provide far more insight on self-presentation attempts than those observed across the board. Therefore, this research will feature a combination of topic modelling as well as word frequency and collocation analysis to better identify strategic self-presentation on dating sites for male users who may perceive their height, education, weight or race to serve as dissuading factors for prospective long-term partners.

REFERENCES

Bruch, E. E., & Newman, M. E. J. (2019). Structure of online dating markets in US cities. Sociological Science, 6, 219-234.

Finkel, E. J., & Eastwick, P. W. (2009). Arbitrary social norms influence sex differences in romantic selectivity. *Psychological Science*, *20*(10), 1290-1295.

Goffman, Erving. (1959). The presentation of self in everyday life. Garden City, N.Y. :Doubleday

Heino, R. D., Ellison, N. B., & Gibbs, J. L. (2010). Relationshopping: Investigating the market metaphor in online dating. *Journal of Social and Personal relationships*, *27*(4), 427-447.

Shishido, Juan, Jaya Narasimhan, and Matar Haller. "Tell Me Something I Don’t Know: Analyzing OkCupid Profiles." (2016).