PERSPECTIVES ON COMPUTATIONAL RESEARCH

FINAL ASSIGNMENT

Strategic Male Self-Representation Using Written Language in Online Dating

Contents

No table of figures entries found.

Abstract

Strategic self-representation assumes centrestage in online dating, but the different approaches undertaken by distinct demographics remains largely unknown. We apply a topic modelling approach that extends the work of focus exclusively on the communication styles of men. We check for differences in the presence of these distributions across the different levels for each of these categorical variables. We use Bruch et al (2019)'s framework to help determine the Using two

different modelling approaches of Latent Dirichlet Allocation (LDA) and Non-Negative. However, we find. We find no clustering of documents

Introduction

'So tell me about yourself'. A commonly asked question in day-to-day interactions, is usually met with silence and hesitation. That can no longer be the case for the 1.67 trillion online dating industry, which is entirely predicated on the exercise of 'getting to know' people and locate prospects for both short and long-term romantic relationships. Self-introduction, and by extension- self-marketing now assume central importance.

Long before the advent of online dating, self-representation had long served as grounds of debate in social anthropology. Goffman (1959) elaborated on the means through which each individual would engage in strategic activities "to convey an impression to others which it is in his interests to convey". Perhaps consequently, he regarded the 'self' as malleable and a work-in-progress constructed *by* the dynamics of each social interaction, rather than driving *them*. This fluidity has only become emphasized with the specific affordances of social media networks that have allowed individuals to build different online 'selves' (Papacharissi, 2010), where some aspects may *be built* beyond their initial intentions.

We may thus think of self presentation- also referred to as impression management- refers to the various processes by which people try to control the impressions that others form of them (Psych Textbook), but may. Goffman further distinguishes between two aspects of self-representation in terms of intent. Firstly, what is consciously 'given'- which social psychologist would later rather to as self-disclosure. Secondly, what in unconsciously 'given off' – including appearance, body language, gaze and props, etc.

One of the most critical avenues for compelling self-presentation may be that of mate selection, which several economists and sociologists have likened this to 'marriage marketplace'. A number of online dating service providers in developed countries may facilitate expansion of potential mates beyond the limits of even extended offline social networks (Piskorski, 2014). Caciopo et al (2013) assert that as many of one in three marriages in the United States is facilitated through these portals- a finding that has been referenced by several players in the industry as evidence of deep, fulfilling relationships on offer.

Self-Concept and Mate Selection

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). These will then be harnessed to address questions of self-

representation in the essay section specifically for male users. Theoretical explanations for the choice of sexual mates span a variety of disciplines from psychology to evolutionary biology. We will adopt one in particular recently highlighted by Bruch et al (2017) using similar data- the triad of opportunity, preferences and opportunity-influenced preferences.

While our findings are intended to validate clearly defined hypothesis, the features that can help with those process are not readily available in the raw data. This confirmatory study will thus rely on methods drawn from exploratory analysis. We will attempt to develop 'groups' of words or communicative choices that can be thus compared against existing demographic variables in the dataset. These will includes abstract clustering of profile essays using vector-based methods, followed by topic modelling supplemented by frequency analysis.

The theory section first explores self-representation generally and then specifically with respect to online dating. This nuanced background will then help assess the avenues for self-representation on dating sites and existing research on their influences. We will then delve into the three possible hypotheses and establish how they will be assessed by means of our chosen data. The data section lays out descriptive statistics of the key variables of interest.

The Elements of Online Self-Representation

Online dating systems are usually hosted on websites that offer three core functionalities- the ability to post personal profiles, view others' profiles (often with reliance on basic demographic and geographic filters or algorithmic recommendation systems), and send private messages to them (Fiore et al, 2017). Self-presentation here typically includes one or more profile pictures, fixed-choice responses (for demographic details such as age, race, education, etc) and at least one free-form text essay for self-introduction. Fiore (2010) found no relation of the fixed choice responses to overall profile attractiveness.

Several studies have explored the independent and relative role of pictures vis-à-vis the rest of the features. Hitsch et al (2010) build a model around attractiveness that depends exclusively on physical appearance in photos. Moreover, Fiore et al (2010) find that photos play the largest role in determining overall profile attractiveness for both men and women. Emory (2017) add that women in the US typically accord 50% higher importance to their suitors' descriptions relative to men (who place greater weight on the photos). Thus, our study will focus exclusively on the text descriptions for men.

For male users, Fiore et al (2010) found no association between the perceived attractiveness of text descriptions with the length of the text itself, use of self-reference, or the proportion of emotionally-charged words. Intriguingly, Nagarajan and Heast (2009) did find such a relationship for text essays rated as feminine. This connects with the claim by Brand et al (2012) that aggression is correlated with masculinity and by extension to testosterone, and is seen as desirable only by women at a certain stage of their ovulation cycle.

Why is the Self-Summary an Arena for Strategy?

Self-representation- whether online or offline- is fraught with three tensions.

Firstly, Mead (1964) views selfhood as an "interplay between one's own judgment and perceived judgment of others" based on memories of past encounters and social expectations placed on its participants. Interactions in online environments impose these same contradictory drives for openness and autonomy as in offline settings (Greene, Derlega, & Mathews, 2006). However, the computer mediated communication bestows greater control over self-presentational behavior, thereby allows individuals to manage their online interactions more strategically. Moreover, the asynchronous nature of such communications and the primacy of verbal and linguistic cues over less controllable nonverbal communication cues imply that online self-presentation is more malleable and subject to self-censorship than face-to-face self-presentation (Walther, 1996). In Goffman's (1959) terms, more expressions of self are "given" rather than "given off."

Secondly, individuals may be torn between authenticity and adaptation. They may state objective reality and projection of aspired versions of themselves. These usually match with what is perceived as more socially desirable- likable, competent and moral (Schlenker (1985, 1986). Alternatively, they may tailor their public images to the perceived preferences of their communicative partners. This adaptive process can simply involve emphasis of different sides of themselves to different audiences while maintaining accuracy and veracity for each one. The extent of self-representational freedom becomes limited with intimates relative to strangers, where the former being more cognizant of the actual self. In some forms of online communication- such as anonymous chatrooms- this could incentivize modified representations. However, in online dating- where users expect to become intimate with prospective partners- the scope for falsehood reduces notably.

Finally, individuals would choose between established roles or collectively building new ones. Online dating has only recently moved from a taboo relegated to social misfits to a widely acceptable practice. , norms around the institution of online dating are still evolving. In the absence of clearly established communicative protocols, its users may turn to more established and pervasive gender roles. Conversation patterns have been notably different- some differences embodied in the impressions the sexes choose to project. For example, men are expected to be boastful and self-assured, images of confidence are valued more for men than women in most cultures (William and Best, 1990). Shishido et al (2016) have already established that the differences in topic choice across genders exist only for the heading of 'enthusiasm' Due to the challenges of extracting such aspects, we will refrain from exploring this tension in this study.

Comment [H1]: Reference

These can be seen in explicit verbal claims, but also more subtly about the topics they choose to talk about and the particular attitudes they express. Attitudes can prove difficult to mine from text data, and could be relegated to future studies.

Theory

The Opportunity Dimension

Opportunity refers to available romantic options within an acceptable distance. For example, in 'thinner' markets- such as rural areas, or for homosexual couples, options are limited relative to heterosexual couples in metropolises. However, opportunity has long been a constraining factor. Across diverse cultures, the selection of mates in heterosexual couples has typically been characterized by homophily (Fiore, 2005) as established by the macrosocial order (Blau, 1977).

Historians such as Coontz (2006) and Karandashev (2015) argue that the selection of marriage partners on the grounds of personal compatibility and romantic love over and above purely economic considerations has been recent in human history, and has emerged in the post-Industrial Revolution era, parallel with the rise of women's earning power. Economists such as Hitsch et al (2010) have represented these dynamics in the form of 'marriage markets'. The tendency to marry within similar groups are attributed to the requirements of work and education and cast as 'frictions' to a freer selection of mates purely on preference. Selection has thus been horizontal (among the same social stratum) rather than vertical (open to all individuals acceptable based on individual preferences).

In the last twenty years, the widespread adoption of dating services providers in developed countries- both online and offline matchmakers- may overcome these frictions and facilitate expansion of potential mates beyond the limits of even extended offline social networks (Piskorski, 2014). Caciopo et al (2013) assert that as many of one in three marriages in the United States is facilitated through these portals- a finding that has been referenced by several players in the industry as evidence of deep, fulfilling relationships on offer.

For the sake of this study then- which focuses on the increased access to prospective romantic partners via online platforms- the opportunity dynamic will not be actively considered.

The Preferences Dimension

Contrary to standard model of rational economic choice, Slovic (1995) points out that preferences are often discovered ex-ante. A number of studies have already explored the dynamics of this process in the context of potential romantic partners meeting in person. Speed dating provides the benefit of greater 'revealed preference' flexibility by allowing every male participant to interact with every female participant. MacFarland and Jurafsky (2013) delineate how vocal projection by men, and interrupting women to demonstrate empathy were highly correlated with selection by the latter.

However, such opportunities afforded by live interaction- whether strategically used or 'given off' in Goffman's sense- are typically unavailable to those using dating apps. The cues

provided to the other party now depend entirely on what is 'given', i.e the users choices to share and conceal in their virtual self-description.

In earlier work, Bruch et al (2015), they suggest that users of dating sites select whom to contact through a three-stage decision process. In the presence of a vast number of options, and given the constraints of time, working memory and computation ability, individuals may rely on more automatic heuristics. At the first stage, screening may be conducted based on either on conjunctive (deal breakers) or disjunctive (deal maker) factors. Given the tendency towards homophily observed above, it would not be surprising for users to shrink the applicant pool by screening on the grounds of ethnicity and education. At a more biological level, female users may also use height as a filter (Toma et al, 2010) or level of physical fitness (reference). These mechanism will be explored extensively in this study.

Heino et al (2010) argue that these preference-driven avenues further entrench the economic dimension of mate selection through an acute, implicit awareness of 'relationshopping'. Herein, potential partners are reduced to entries in a catalogue to be scrolled through. In this sense, they suggest an emerging conscientiousness of 'marketing', with the product being themselves, and the potential mate assuming the role of a buyer (ibid). This perception thus links the private worlds of romantic intimacy with those of mass consumption and broader perceived appeal to the opposite sex. In this vein, male users could either be catering to highly specific niches, or to the mass market respectively. These two more commercial interpretations tie in directly with Bruch (2015) et al's exploration of aspects of evolutionary psychology underlying preferences.

Under the **matching hypothesis**, mate preferences could be highly idiosyncratic and specific. In our dataset, this would imply that we would see no clear pattern to what is being discussed. We would expect no dominant topic or group of topics to be referenced repeatedly. Alternatively, under the **competition hypothesis**, consensus already holds on what constitutes an attractive profile, as it does across a range of species (ibid).

With respect to physical qualities, like height, this perception does exist. However, we are yet to determine through these data whether certain personality traits fall within this consensus too. We have already failed to find evidence on gender roles being represented. There may, however, be a pattern to the topics under discussion, and what they 'give off' about the user. In this study, we test these two competing hypotheses by checking for variables that may be considered potentially important to the selection of prospective male partners in offline settings- height, fitness level, race and education.

The Opportunity and Preferences Dimension

Bruch et al (2015) emphasize that the two dimensions of opportunity and preferences are not mutually exclusive. The perceived preferences of an individual may be shaped by what is repeatedly observed among potential mates. In this sense, homophily may be seen purely as

a reaction to limited exposure to groups distinct in their ethnic or educational background, rather than a clearly defined preference. Under this condition, we would expect that 'disadvantaged' male users would see no need to adopt 'compensatory' strategies. They would simply represent themselves as a novel experience, and market themselves accordingly.

In the context of online dating, the status quo 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 data. Nonetheless, we will be able to test which subgroups of individuals aim to either confirm or deviate from these norms. Whether this impression is 'given' consciously or unconsciously 'given off'- perhaps as an extension of Heino' (2010)'s self marketing- or not can only be confirmed through subsequent interviews.

To construct this sense of norm, we look into the ongoing joint distribution of topics and find the most common patterns observed. We then spot distributions of topics away from this guideline.

DATA & METHODS

Data

Overview

The data set used for this study is publicly available. 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. 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 18330.

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.

This data includes 10 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.

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.

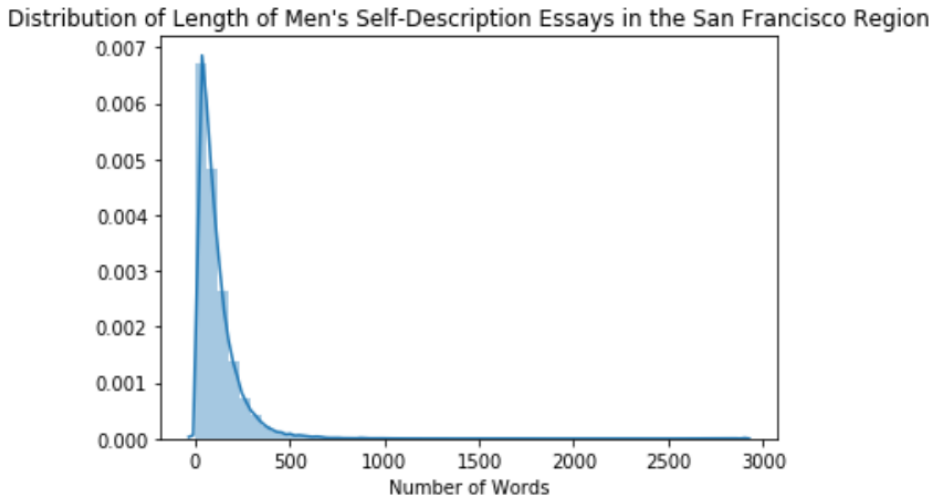
Descriptive Statistics

Extending Goffman's (1959) classification, we address variables as those that are explicitly stated with the intent of self-disclosure, and those that may be inferred and the deliberateness is less certain.

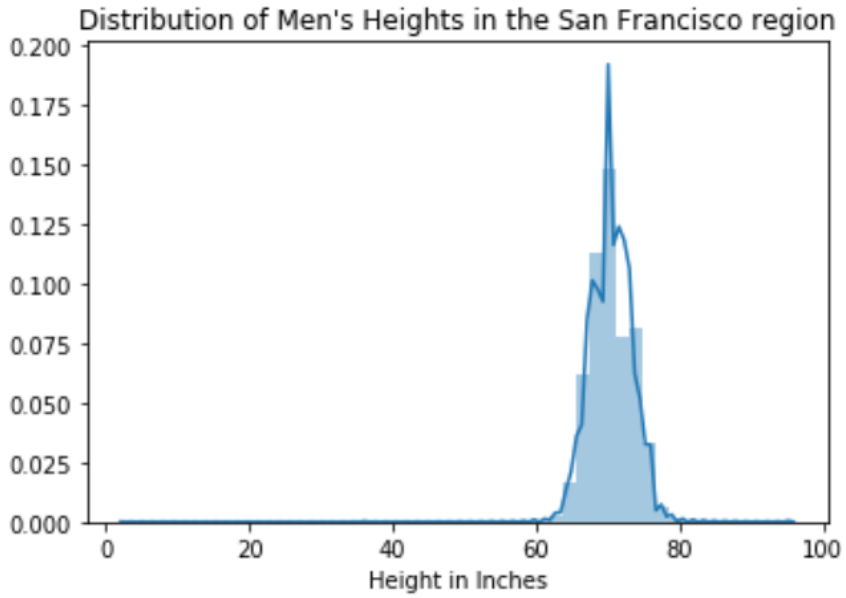
Given-Off Variables

Essay Length

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.



Height



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

HEIGHT

Female users of dating sites cited height as an important element of their preferences for male partners (Shepperd & Strattman, 1989). Interestingly, height did not feature as an important factor in rating attractiveness of their partners post-hoc in real world couplings (ibid). This fits with the Slozic's revealed preference hypothesis.

Nonetheless, OkCupid (2010) found for both American men and women that median height lay two inches above that of the United States, suggesting over-representation. Men demonstrated a tendency to inflate number closer to the benchmark of 'six feet'. These trends highlight a consensus- in line with the competition hypothesis- that this particular trait is fundamental to the attractiveness of a potential partner. Instead of using an absolute

value, we rely on quantiles- and establish the 25th percentile as the cutoff for being classified as short.

The statistics for the height variable confirm our hunch from the visualization that height is largely normally distributed, while essay size is heavily skewed. There seem to be outliers and incorrectly entered data on the lower end of the range. However, these data points were important for the core questions of this research around their answers to questions on text analysis. Hence, they have been retained for now.

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.

EDUCATION

Education has been considered an important element of the attractiveness of mates, often in different directions based on gender. Preference for women's education stays limited in the United States to a bachelor's degree, and falls sharply thereafter (Bruch et al, 2019). The preference for highly educated men as partners has been observed widely (Stevens & Schaefer, 1990), potentially due to correlations with intelligence and earning power (Fiore et al, 2015). Less educated male users on the site may adopt strategies to gloss over this perceived shortcoming.

The vast number of terms were grouped using the High School Diploma as the cut-off point

FITNESS

Following the framework used in the University of Michigan Computational Social Science workshop, we group

The following classifications were used:

Finally, physical fitness may be linked to evolutionary biology in terms of men's ability to provide, as well as a proxy for a woman's fertility (Fiore et al 2017). While no specific research has been conducted on this aspect per se, several media sources () refer to men's

tendencies to display their physical strength through shirtless 'selfies'. Regardless of whether this lies within the desired traits for female users, it does seem to be perceived as desirable by the male users themselves.

RACE

Race has been a contentious issue, as the United States evolves from a biracial to multiracial society. Lin and Lindquist (2013) provide a comprehensive review of the racial dynamics of interactions initiated and sustained across users from different backgrounds. They leverage data from an unidentified online dating platform and discover preferences of racial homophily across male and female participants. In particular, women were found to respond to messages sent by men belonging to a group with equal or higher perceived racial status. Over 50% of respondents in this OkCupid dataset identify as white, with a number of mixed ethnicities.

If the preferences dimension were considered to be dominant, we would see the same set of topics and words appearing in all profiles, regardless of specific categories under the demographic variables outlined above.

METHODS

Preprocessing

First, we remove any HTML tags and or special characters. We also correct any recurring spelling mistakes, which occur primarily through accidentally compounded words. While the number of mistakes could have itself served as an interesting aspect of the analysis, we do find that the writing has been largely grammatically correct.

Tokenization

The cleaning process reveals the presence of over 10,000 unique tokens. Including bigrams and tokens with frequency of at least greater than one raises this figure to

*The use of slang terms stays the same across the board

We also compute the complexity of these messages through the Flesch Kincaid level, the proportion of words with three or more syllables.

User Type	Profile Length (in words)	Flesch Kincaid Reading Level	Proportion of Long Words
General Male User	80	6.73	0.20

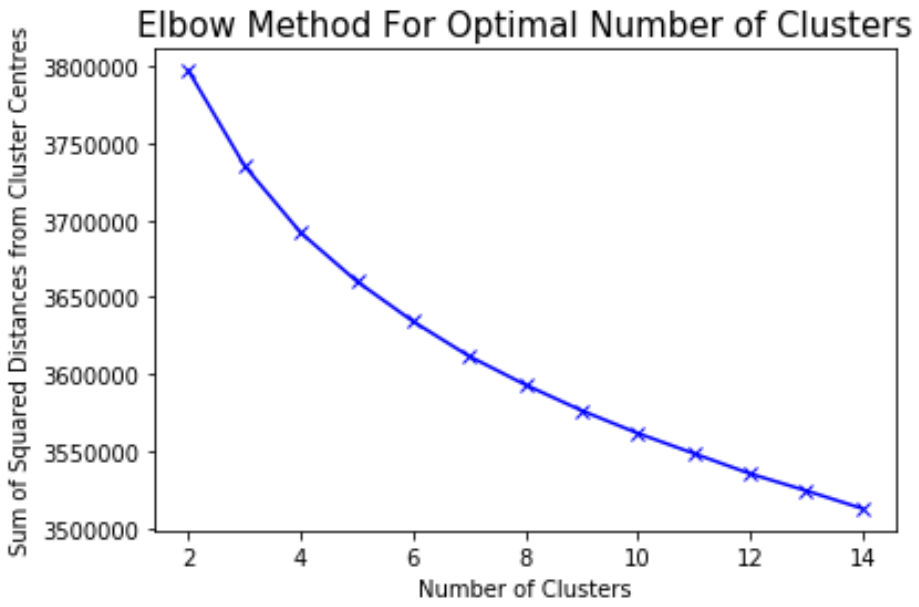
Users with more than a High School Diploma	85	7.1	0.21
Users with a High School Diploma or less	66	5.63	0.18

Abstract Representation

We can also represent each document as a vector, under the Doc2Vec algorithm within the Gensim library in Python. Under this method, all the word vectors within each document are averaged to get one vector score for the document as a whole. Then documents can be compared for similarity based on this total score.

Raising the size of the vector for the documents reduces the set of 'similar' dating profiles on average for each profile. The extent of similarity (measured through cosine similarity) continues to fall.

We apply Principal Components analysis to all these dimensions and then aim to identify clusters. For this, we rely on the k-means algorithm (based on what work). However, t



With both a regular plot of the PCA components, as well as t-SNE, we see a similar output. : a near elliptical scatter of data points, there are no discernible clusters of documents. We do not pursue this abstract vector-based approach any further, and turn instead to the more human-understandable realm of topic models.

Data Cleaning

The entire study runs on two parallel tracks- analysis of the core 'About Me' essay in the dating profiles, and then comparing results based on key demographic variables.

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.

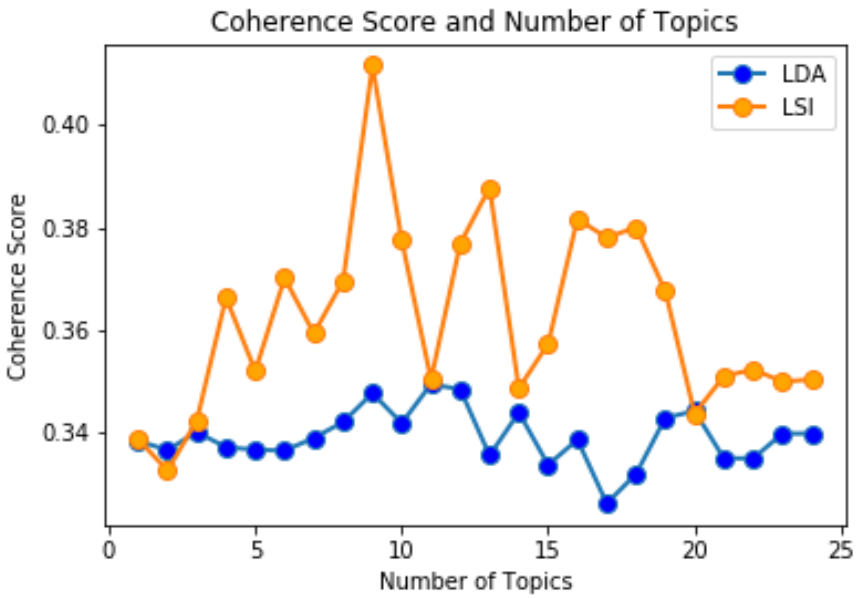
Topic Model Selection

We apply 4 algorithms to test for the ideal number of topics- Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSI), Hierarchical Dirichlet Process (HDP) and NMF NMF (Non-NegativeMatrix Factorization). The first three rely on Term Frequency Matrices, while the third also incorporates the Term-Frequency Inverse-Document Frequency matrix.

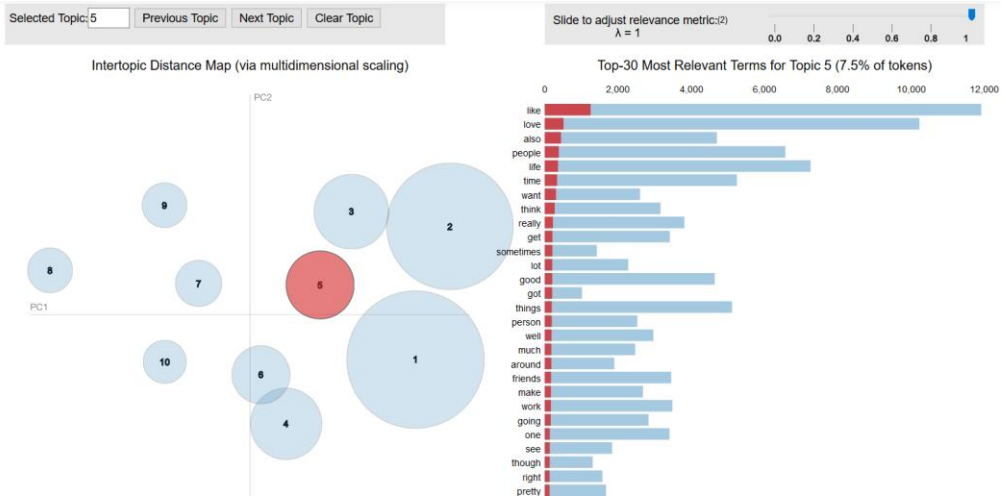
HDP automatically selects the required number of topics at 20. However, an inspection of the words appearing under each heading reveals a large overlap in the words used within each topic. Hence, this approach is unlikely to be suitable for our objectives.

The final choice of the correct number of topics draws on both qualitative and quantitative approaches. Under quantitative, we check for the coherence of the topics generated.

LDA and LSI reveal that the highest coherence scores occur for 10 and 8 topics respectively, as shown below.



By plotting these topics on the PyLDAvis, we get clear separation of topics (no overlaps in the bubble plot) and a large size for each. However, on expanding the top number of words to be displayed under each topic, the topic itself is not entirely clear from the words included at a more intuitive, human level.



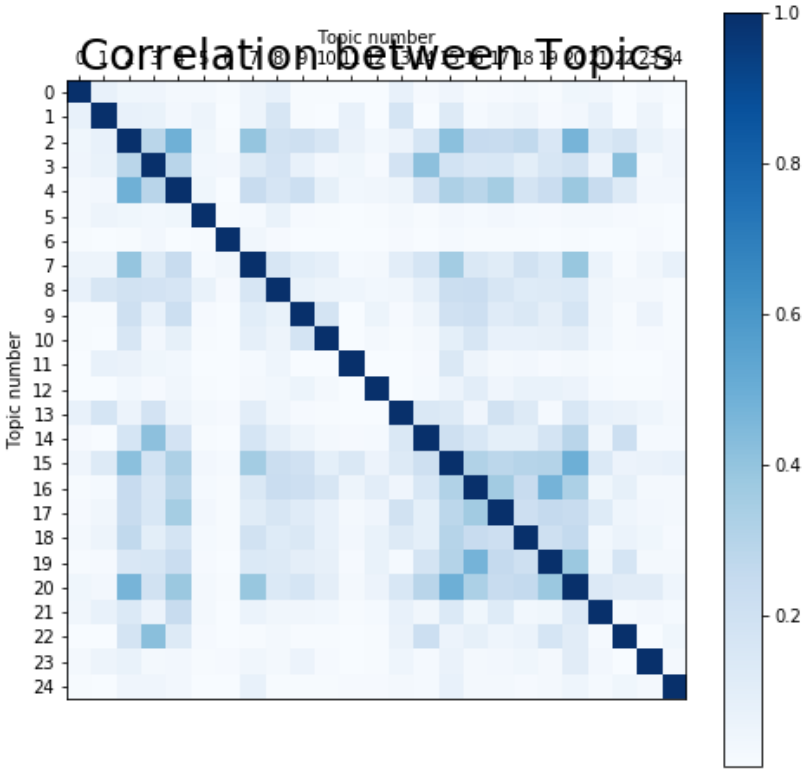
This leads to differing recommendations on the ideal number of topics, with 20 sample words. The output from each is provided below. We also explore the pyLDAvis visualization of the topics themselves to check for visualizations.

It may be argued that truly distinct topics would show minimal overlap in the circles on the bubble plot. However, the specific nature of the topic under discussion- being related to subjective personal self-presentation- may allow for overlap among.

Even for LDA, some of the 10 topics provide little scope for interpretability due to being too general. In this sense, the 'optimal' figure of 8 topics under LSI seems far too low for our objectives.

Hence, we inspect the result of the LDA model more closely. The topics are shown to overlap to some degree.

For this purpose, we draw largely on the framework of Shishido et al (2016) in terms of interpreting the topics. Finally, in spite of coherence measures indicating otherwise, we find that the most 'intuitive' and interpretable set of topics appear fro r**25 topics under the NMF model.** We cross compare the



The NMF (Non-negative Matrix Factorization) approach relies of TF-IDF (Term Frequency and Inverse Document Frequency) Matrices, rather than the pure Term Frequency matrix for LDA. This suits the situation we are operating in. NMF is faster than LDA and often works a little better for small documents like those in the current situation.

With these topics selected,

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.

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."

Topic Selection

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.

Topic Visualization

We also check for the similarity of the topics to each other. This follows the approach used by the University of Michigan.

A cursory glance through the most popular topics (comprising the highest proportion on average) reveals the following headings:

Novelty
Carpe Diem
Fun Loving

This seems to fit with the notion of the expected social role.

Topic Clustering

Having established the most coherent distribution of topics, we then plot the topics themselves along principal components to understand the clustering of the core ideas. Using pyLDA visualization reveals four major clusters of topics.

Topic Distribution

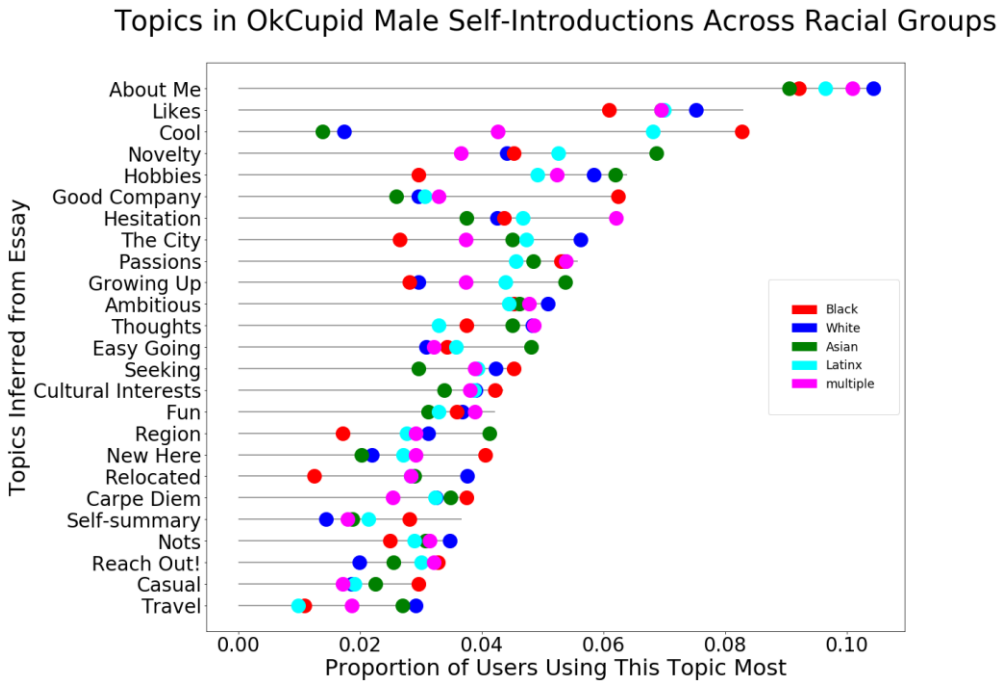
We arrange the most popular topics- which feature in the largest number of profiles.
We try to understand what words each topic has in common with the others, and different from others.

In our diagrams, we see a correlations between ttopics

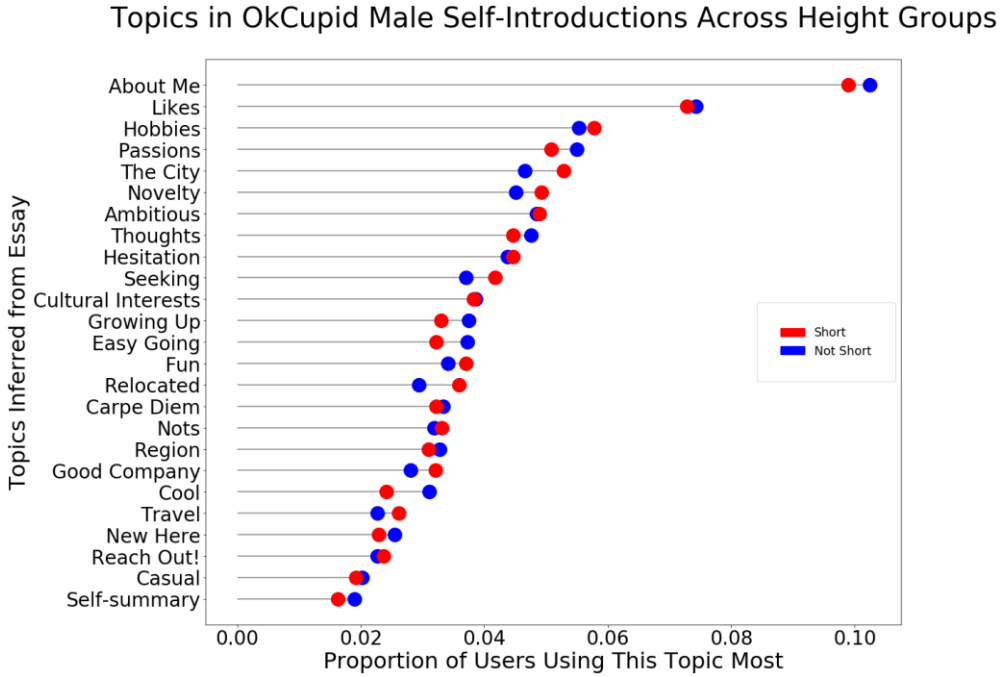
We will then aim to check which are the most common combinations of topics in this corpus. This will form the 'Status Quo' as per the theoretical section.

Results

In this section, we consider the mean proportion of profiles composed by the words of a given topic, and compare them for each sub-group.

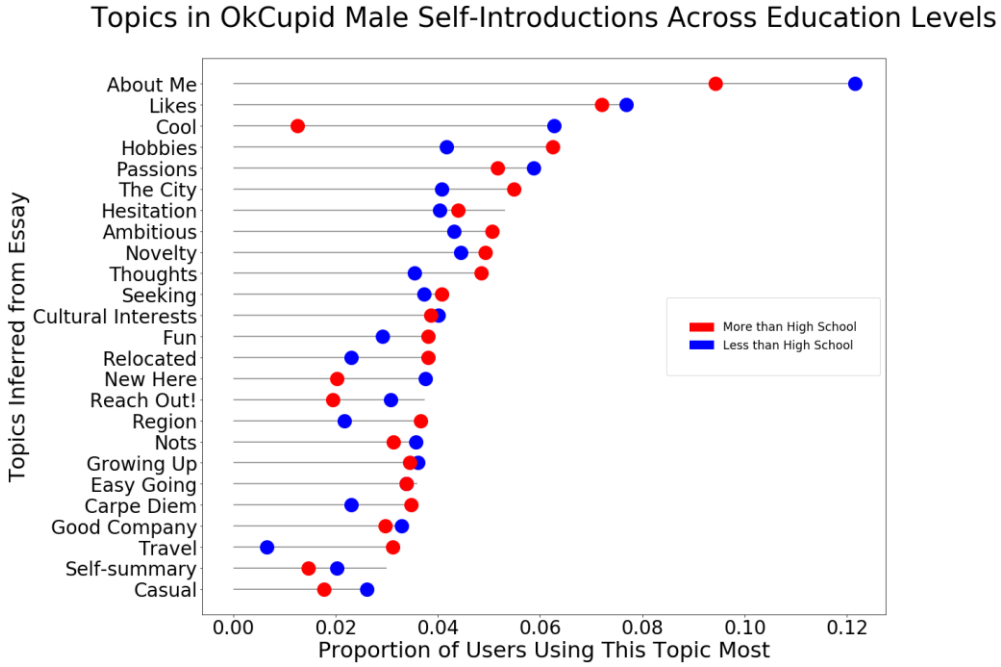


In the category of race, we again see similar patterns. There are a few interesting patterns- such as the high use of words to do with 'cool' by African American men, and a relative lack thereof in Asian and White men. Black men also trail behind in their references to relocation and the region, and yet use words from the topics on being new. This may imply that they wish to emphasize being new and on the lookout for friends without mentioning where they are from, which might carry the burden of judgment.



There are almost no visible differences between the average representation of each topic

EDUCATION



We observe some differences in a few topics, such as 'cool' and 'travel' holding higher proportions for those with less and more than high school diplomas respectively. In other areas such as 'Reach Out' and 'New Here', we observe only a few percentage points' difference.

0

12- Hobbies

1- Relocated

7- Enthusiasm (Love)

11- Carpe Diem

6- Likes

12

FITNESS

Favoured Topics:

Fit

```
'topic_0', 'topic_12', 'topic_1', 'topic_7', 'topic_11', 'topic_6',
'topic_2']
```

```
Not Fit
['topic_0', 'topic_12', 'topic_6', 'topic_1', 'topic_7', 'topic_2',
'topic_11']
```

EDUCATION

High School or Less

```
['topic_0',
 'topic_12',
 'topic_7',
 'topic_6',
 'topic_2',
 'topic_11',
 'topic_13']
```

More than High School

```
['topic_0', 'topic_12', 'topic_1', 'topic_6', 'topic_7', 'topic_11',
'topic_2']
```

HEIGHT

Short

```
['topic_0', 'topic_12', 'topic_6', 'topic_1', 'topic_7', 'topic_11',
'topic_2']
```

Not Short

```
['topic_0', 'topic_12', 'topic_1', 'topic_7', 'topic_6', 'topic_2',
'topic_11']
```

RACE

White

```
['topic_0', 'topic_12', 'topic_1', 'topic_6', 'topic_7', 'topic_2',
'topic_11']
```

Black

```
['topic_0',
 'topic_12',
 'topic_6',
 'topic_7',
 'topic_11',
 'topic_13',
 'topic_14']
```

Asian

```
['topic_0', 'topic_12', 'topic_2', 'topic_6', 'topic_1', 'topic_11',
'topic_7']
```

Multiple

```
['topic_0', 'topic_12', 'topic_7', 'topic_1', 'topic_11', 'topic_6',
'topic_2']
```

LatinX

```
['topic_0', 'topic_12', 'topic_7', 'topic_6', 'topic_1', 'topic_2',
'topic_11']
```

NMF Analysis- Top Requirement

LDA ANALYSIS

For the LDA topics, almost identical

Education
Less- ['topic_21', 'topic_3', 'topic_18', 'topic_10',
 'topic_16',
 'topic_20',
 'topic_15']
More than High School
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_15',
 'topic_10',
 'topic_20']

Height

Short
['topic_21',
 'topic_3',
 'topic_18',
 'topic_15',
 'topic_16',
 'topic_10',
 'topic_20']

Not Short
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_15',
 'topic_10',
 'topic_20']

Fitness

FIT
['topic_21',
 'topic_3',
 'topic_18',
 'topic_15',
 'topic_16',
 'topic_10',
 'topic_20']

NOT FIT
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_10',
 'topic_15',
 'topic_20']

Race

White

```
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_15',
 'topic_10',
 'topic_20']
```

Black

```
['topic_21',
 'topic_3',
 'topic_18',
 'topic_15',
 'topic_16',
 'topic_10',
 'topic_12']
```

Asian

```
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_15',
 'topic_10',
 'topic_20']
```

Multiple

```
['topic_21',
 'topic_3',
 'topic_18',
 'topic_10',
 'topic_16',
 'topic_15',
 'topic_20']
```

LatinX

```
['topic_21',
 'topic_3',
 'topic_18',
 'topic_16',
 'topic_10',
 'topic_15',
 'topic_20']
```

Conclusion

The essay section of online dating websites provides users with opportunities for strategic self-representation towards securing romantic partners beyond the limits imposed by offline social networks and disadvantages involved with physical and educational characteristics. We chose to

explore these dynamics specifically for male heterosexual users in the US through the scraped profiles of about 18,000 users. Given the considerable limitations imposed by the data- such as the absence of photos or measures of 'success' in terms of subsequent online and offline interactions, our research has still managed to unearth valuable insights.

Though the essays themselves do not cluster in any mathematically recognizable format, representation through self-disclosure becomes available through topic modelling. We find that through two different strategies- Latent Dirichlet Allocation and Non Negative Matrix Factorization- and by analyzing them against subsets of 4 key variables- weight, height, education and race- that tendencies exist both towards a set of core topics considered collectively attractive as well as high variation in supplementary topics. The first model shows almost no variation in the top 7 topics, while the second shows reordering and occasional introduction of previously unseen topics for . As for proportions of users who have the highest proportion of their profile based on one topic, there are no significant differences across various demographic variables.

Within its own design, the study suffers from certain limitations. Not all aspects of self-presentation would necessarily be included in this first 'About Me' essay. As seen in Shishido et al (2016), an analysis of other sections (such as hobbies, music and TV shows) may reveal further pattern in self-disclosure. Nevertheless, it may be argued that many of these entries do not provide the necessary freedom for free linguistic expression. Many could be reduced to fixed response entries, or filled by importing 'likes' on Facebook or specific preferences via other through integrations with other apps such as Spotify for musical tastes. Conversely, many competitor data apps such as Tinder and Bumble only require one essay. From this lens, the approach may actually prove more valid for future studies (though reduce some of the relevance for his study itself).

Furthermore, the generalizability of these findings may be questioned in terms of user psychographics and cultural nuances. The profiles were specifically taken from the San Francisco region- known to be home to Silicon Valley, and a higher degree of comfort with technology and digital platforms. Besides the requirement of comfort with technology in general, OkCupid's requirement of writing extensive essays for their profiles creates a further selection bias towards more self-aware and patient users. Therefore, the behavioural inferences from this platform may not be as directly applicable to users of other popular platforms like Tinder, which focus on quick swiping-based selection and limited text. They may also not extend directly to non-occidental cultures in India and China where dating has experienced notable growth, but may be harnessed in markedly different ways.

Comment [H2]: Add reference

Finally, the passage of time since the collection of these data could render some of the inferences less relevant. Like in Bruch's most recent work (2019), male mating behaviour often involves checking for what confers success to competitors. They find that as many as 5% of visitors to mens' profiles are other heterosexual men. This marks a trend to assess what the 'status quo' is, and how to help their profiles stand out. The 'relationshopping' awareness may be more deeply entrenched now than had been the case in 2012. Thus, follow-up studies done on OkCupid Data from the same region, or even other cities, may provide insight into whether this insight has grown. Without data on the eventual online matches and offline meetings from these dates, we do not know which profiles were selected more often/ rated as more attractive. Without follow-up interviews or, it is impossible to measure whether the specific choice of words was aimed at authenticity, or matches

with an awareness of 'relationshopping'. The latter could have potentially been inferred by the date of creation of the profile, with longer-term users perhaps gaining more 'experience' and conscientiousness of what achieves their self-designated 'dating success'. Future studies could pair topic modelling with messaging outcomes. For example, Bruch et al (2019) touch on similar data for San Francisco.

Comment [H3]: Could be slightly trimmed

Finally, within the realm of text analysis, new advances could allow for checking more aspects of 'given-off' cues like confidence. Brand et al (2012) uncovered strong correlations between attractiveness of photos and text components of men when viewed independently by women so as to prevent a spillover halo effect (Dion & Waldster, 1972). They hypothesize that the confidence observed in the text served as the mediating variable. This confidence, in turn, may result from habituation to receiving the appreciation of the opposite sex, which in turn may be linked to their physical attractiveness. However, in the absence of photos and given the limited direct relevance of fixed choice options, confidence would need to be determined through alternative means.

Fortunately, the OkCupid dataset provides a potential avenue to extract this feature from the text itself. Selectivity about mates has typically been observed more in women than men (Finkel et al, 2011, Fiore et al, 2017). The final essay, entitled 'You should message me if...' helps communicate selectivity on the part of the males, and reverse the 'relationshopping' paradigm. Some profiles reply with statements to the effect of automatically approving any woman who approves of them. Others lay down lists- often long and detailed- of attributes that the men seek in women. Such high expectations could act as a signal- either 'given' off or 'given'- of self-awareness, high demand by and experience with other women as well as the ability to choose. These cues could collectively proxy for confidence even in the absence of any photographs. The responses for this essay have already been cleaned, and could easily be leveraged for a follow-up study combining topic modelling and inferred levels of confidence among male users.

Conclusion

Bibliography

There are no sources in the current document.

REFERENCES

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

Bruch, E. E., & Newman, M. E. J. (2018). Aspirational pursuit of mates in online dating markets. Science Advances, 4(8), eaap9815.

Bruch, Elizabeth, Fred Feinberg, and Kee Yeun Lee. 2016. "Extracting multistage screening rules from online dating activity data."Proceedings of the National Academy of Sciences of the United States of America113: 10530-10535

Bruning, R. H., Schraw, G. J., & Ronning, R. R. (1999). *Cognitive psychology and instruction*. Prentice-Hall, Inc., One Lake Street, Upper Saddle River, NJ 07458.

Cacioppo, J. T., Cacioppo, S., Gonzaga, G. C., Ogburn, E. L., & VanderWeele, T. J. (2013). Marital satisfaction and break-ups differ across on-line and off-line meeting venues. *Proceedings of the National Academy of Sciences*, *110*(25), 10135-10140. Coontz, S. (2006). *Marriage, a history: How love conquered marriage*. Penguin.

Chappetta, Kelsey C. , Barth, Joan M., 'How gender role stereotypes affect attraction in an online dating scenario,' Computers in Human Behavior, Volume 63, 2016, Pages 738-746, ISSN 0747-5632,

Curington, C. V., Lin, K. H., & Lundquist, J. H. (2015). Positioning multiraciality in cyberspace: Treatment of multiracial daters in an online dating website. American Sociological Review, 80(4), 764-788.

Dion, K., Bersheid, E., & Walster, E. (1972). What is beautiful is good. Journal of Personality and Social Psychology, 24, 285–290. doi:10.1037/h0033731.

Duck, S. (2000). *The social psychology of personal relationships* (Vol. 3). John Wiley & Sons Inc.

Eastwick, P. W., & Hunt, L. L. (2014). Relational mate value: Consensus and uniqueness in romantic evaluations. Journal of Personality and Social Psychology, 106(5), 728.

Emory, Leah(2017), Bustle. 'How Many People Who Meet On Dating Apps Get Married? Swiping Isn't Just For Hookups', Retrieved from - https://www.bustle.com/p/how-many-people-who-meet-on-dating-apps-get-married-swiping-isnt-just-for-hookups-44359

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

Fiore, A. T., & Donath, J. S. (2005, April). Homophily in online dating: when do you like someone like yourself?. In CHI'05 Extended Abstracts on Human Factors in Computing Systems (pp. 1371-1374). ACM.

Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. (2008, April). Assessing attractiveness in online dating profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 797-806). ACM.

Fiore, A. T., Taylor, L. S., Zhong, X., Mendelsohn, G. A., & Cheshire, C. (2010, January). Who's right and who writes: People, profiles, contacts, and replies in online dating. In *2010 43rd Hawaii International Conference on System Sciences* (pp. 1-10). IEEE.

Gibbs, J. L., Ellison, N. B., & Heino, R. D. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in Internet dating. *Communication Research*, **33**(2), 1–26.

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.

Herek G. M. (1986) On heterosexual masculinity: Some psychical consequences of the social construction of gender and sexuality. American Behavioral Scientist 29(5): 563–577. doi:10.1177/000276486029005005.

Hitsch G, Hortaçsu A, Ariely D (2010) Matching and sorting in online dating. Am Econ Rev 100(1):130–163.

Lin, K. H., & Lundquist, J. (2013). Mate selection in cyberspace: The intersection of race, gender, and education. *American Journal of Sociology*, *119*(1), 183-215.

Karandashev, V. (2015). A cultural perspective on romantic love. *Online Readings in Psychology and Culture*, *5*(4), 2.

Kim, Albert Y. & Escobedo-Land , Adriana (2015) OkCupid Data for Introductory Statistics and Data Science Courses, Journal of Statistics Education, 23:2, , DOI: 10.1080/10691898.2015.11889737

Kinch, J. W. (1963). A formalized theory of the self-concept. American Journal of Sociology, 68(4), 481-486.

Leary, M. R., & Tangney, J. P. (Eds.). (2011). *Handbook of self and identity*. Guilford Press.

Lundy, D. E., Tan, J., & Cunningham, M. R. (1998). Heterosexual romantic preferences: The importance of humor and physical attractiveness for different types of relationships. Personal Relationships, 5, 311–325.

McFarland, Daniel A., Dan Jurafsky, and Craig Rawlings. "Making the connection: Social bonding in courtship situations." *American journal of sociology* 118.6 (2013): 1596-1649.

Mead, G. H. (1934). *Mind, self and society* (Vol. 111). University of Chicago Press.: Chicago

Nagarajan, M., & Hearst, M. A. (2009, March). An examination of language use in online dating profiles. In Third International AAAI Conference on Weblogs and Social Media.

OkCupid, 2010, 'The Lines People Tell Online', Okcupid.com, Retrieved from: https://theblog.okcupid.com/the-big-lies-people-tell-in-online-dating-a9e3990d6ae2

Papacharissi, Z. (Ed.). (2018). *A Networked Self and Love*. Routledge.

Piskorski, M. J. (2014). *A social strategy: How we profit from social media*. Princeton University Press.

Shepperd, J. A., & Strathman, A. J. (1989). Attractiveness and height: The role of stature in dating preference, frequency of dating, and perceptions of attractiveness. *Personality and Social Psychology Bulletin*, *15*(4), 617-627.

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

Slovic, P (1997)- The Construction of Preference. American Psychologist

Stevens, G., Owens, D., & Schaefer, E. C. (1990). Education and attractiveness in marriage choices. *Social Psychology Quarterly*, 62-70.

Taylor, D., & Altman, I. (1987). Communication in interpersonal relationships: Social penetration processes. In M. E. Roloff & G. R. Miller (Eds.), Interpersonal Processes: New Directions in Communication Research (pp. 257–277). Thousand Oaks, CA: Sage

Theodore R. Sarbin, "Role Theory," in Gardner Lindzey (ed.), Handbook of Social Psychology (Cambridge, Mass.: Addison-Wesley Publishing Co., 1954), 223-58.

Toma, C. L., & Hancock, J. T. (2010). Looks and Lies: The Role of Physical Attractiveness in Online Dating Self-Presentation and Deception. Communication Research, 37(3), 335–351. https://doi.org/10.1177/0093650209356437