Escort Ratings

ABSTRACT

This paper provides insight into the world of sexual escorts within Brazil. Using a bipartite network of sex buyers (male) and escorts (female), the authors were able to draw connections and a deeper assessment of the sex-trade industry. The data represents both males and females as nodes, with edges forming when the buyer leaves a review for their female counterpart. The dataset, analysis, and graphs have been provided within the paper. Although the data is more explicit in nature, it depicts whether or not the sexual encounter included anal, oral, or kissing, it serves as a backdrop to break down larger real world problems such as viral STD/STI outbreaks.

Brazil was chosen in particular due to its emergence as a growing world power (holding both the last World Cup and Olympics) and their neglect to tackle issues relating to prostitution and human trafficking. The goal of this paper is to provide insight into this often-overlooked social issue. Through the experiments conducted, the overall acumen on virus detection and viral outbreaks can be increased, which in turn would help bring light to controlling this epidemic.

1. Introduction

Brazil has the 13th highest population of individuals with HIV/AIDS and has the 17th largest HIV-related deaths in the world [3]. Outside of Africa or Asia, the highest rate of infection can be seen in Brazil. Also, correlated but not causation, Brazil has legalized prostitution as a profession. This means that sex workers are entitled to social security and other work related benefits [1]. Despite the government's acceptance of the profession, due to socio-economic constraints, HIV and sexually transmitted diseases are still prevalent amongst working females. This is important given Brazil's high rate of newly infected individuals, especially in certain geographic regions and within younger demographics. Although the government is aware of said issues and trends, and their established programs and funding to solve this social problem, the funding isn't being properly allocated enough to solve this issue.

In our paper, we hope to discover trends within the sex trade itself and analyze what conditions would bring about a viral outbreak spreading throughout an entire community. Given our dataset is bipartite, where the edges contain weights and additional attributes, we will begin our assessment on predictive analytics. Based on reviews, who is a buyer likely to choose? Provided a user's preference, who would be the recommended escort? If the buyer is an established member of the social network, what is the likelihood they will reach out to the same escort? We will then take a look at the escort side of the network. Is there a correlation between the escorts responsiveness and the number of encounters she books? Given any two escorts, how much of the clientele is shared?

Both studies will help set the backdrop for getting a better understanding of our dataset. We plan on using this insight to answer, what is the likelihood of a sexually transmitted infection (STI) or sexually transmitted disease (STD) outbreak. We will use greedy algorithms, similar to viral marketing strategies to identify the number of initially infected targets it would take for the entire dataset to get infected. In doing so we will examine an individual's reach, if a buyer or seller were to get infected, to what extent will the STI/STD spread throughout the network? Using the number of days, the seller posts the listing to the time the buyer writes the review, we can also estimate the time it would take for a viral outbreak to spread.

2. Background/Related Work

While on a national level, the HIV and AIDS outbreak in Brazil stabled to .6% of the total population, the prevalence of the issue remains when we step back and examine the outbreak rates against the rest of Latin America. Brazil accounts for 40% of all new infections in Latin America, and while in part this may be true due to its larger population size, it is still one of 15 countries in the world that account for a majority of the number of people living with HIV [1]. The largest rate of infection has been seen among individuals aged 30-49, with a rising trend among the youth. There is also a geographic variance in the infection rates, if the dataset wasn't anonymized, it would be an interesting perspective to see if escorts were from regions, like the south and southeast, where the virus has seen a greater pervasiveness.

Among sex workers, like the ones identified in our dataset, HIV has been found in 4.9% [1]. However, if we look at all sexually transmitted diseases, this number increases to 71.6% having at least one disease, with the most common being: HPV (67.7%) and Chlamydia (20.5) [2]. Now while this study might be a based on a limited sample size, similar studies have shown roughly two-thirds of female sex workers exhibiting signs for at least one disease. These numbers, 4.9%, and 66% will be used in this paper to track the viral outbreaks among our dataset.

While the future of the viral outbreaks within Brazil remains unclear, it is clear that the Brazilian government is taking steps to address this issue. The federal budget incorporates funding to HIV response, which has led to early diagnosis and treatment. However, due to corruption, only around 5% of the budget was properly allocated to "effective prevention", leading to insignificant progress in HIV prevention.

3. Approach

Our approach started off by filtering down the network to a smaller dataset. We started off by removing nodes that had a degree of one or less. Next, we added the giant component filter. We used this component to start our analysis.

The diagram below shows the component in a double circular layout, which strongly complements the bipartite network. The inner circle is the Female Escorts that are colored Pink. The outer circle is Male Buyers that are colored Blue. Both circles are ordered in counterclockwise by the highest number of degrees to the lowest.

Figure 1. Filtered Component Network

We next wanted to perform network analysis to get the base statistics on the giant component.

Table 1. Giant Component Statistics

Attribute	Value
Number of Nodes	9148
Number of Edges	31878
Average Degree	6.9694
Number of Connected Components	1
Bipartite	True
Directed	False
Connected	True
Density	0.000761932023273
Size	31878
Average Shortest Path	5.18033638876
Diameter	15
Average Clustering Coefficient	0.0
Top Degree Centrality Node	(4569, 0.028096643708319668)
Bottom Degree Centrality Node	(16656, 0.0001093254619000765 2)
Top Degree Betweenness Centrality Node	(4585, 0.043413136800205705)
Bottom Degree Betweenness Centrality Node	(16656, 0.0)
Top Degree Closeness Centrality Node	(4585, 0.28389199255121045)
Bottom Degree Closeness Centrality Node	(16585, 0.08652918361555198)
Top Pagerank Node	(4569, 0.0029357206186875393)
Bottom Pagerank Node	(7916, 2.561317676457188e-05)
Top Hub Node	(370, 0.004429594841681599)
Bottom Hub Node	(16585, 2.890336027697638e-15)
Top Authority Node	(4569, 0.009098827681201287)
Bottom Authority Node	(16585, 1.8044436295858464e-15)

Using Gephi, the next step was to analyze the communities

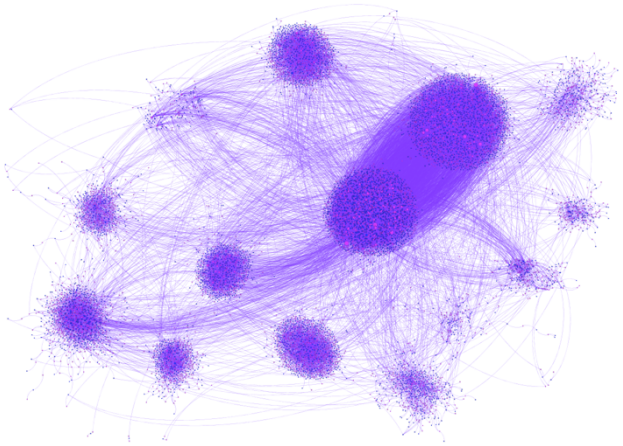

Figure 2. Communities

The graph shows the 17 communities, the blue nodes are the male buyers, and the pink nodes are the female sellers. The analysis parameters and results are shown below.

Table 2. Modularity Analysis

Parameters	
Randomize	On
Use Edge Weights	Off
Resolution	1.0
Results	
Modularity	0.661
Modularity with Resolution	0.661
Number of Communities	17

The basic network analysis provided a high overview of the network and a pivoting point to start to experiment with the data and search for insights. The dataset will be able to answer some of our questions, but some are not fit for the data.

If the buyer is an established member of the social network, what is the likelihood they will reach out to the same escort? Based on our limited dataset, we had no way to determine if a customer

would reach out to the same escort. The data only shows a single interaction between a buyer and seller, if the dataset contained recurring encounters, then we would use action analytics to determine the results. Is there a correlation between the escorts responsiveness and the number of encounters she books? The dataset does not contain proper timestamps and instead give a numeric value of posting days, but not useful to answer this question.

4. Experiment

4.1 Questions

Our experiments will try to answer some of our question we posed in the introductions. Using action analytics, we will use collaborative filtering methods to determine based on reviews and user preference, who is a buyer likely to choose or recommend a compatible partner.

Using network analysis, we attempted to simulate a viral outbreak. Using this simulation, we will attempt to answer the following: what is the growth rate for several different sexually transmitted diseases/infections, what is the effect of centrality on a nodes viral outbreak potential, what is the effect of condoms or other forms of contraceptives being used, and which would decrease the probability of the disease spreading.

4.2 Dataset

For our study, we will be using anonymized data from a public online forum where the male (buyer) who grades their sexual encounters with a female escort (seller). The data was collected from September 2002 - October 2008 and from over twelve cities in Brazil. The dataset was obtained from Konnect [4].

The network has 6,624 escorts, 10,106 buyers, which adds up the total to 16,730 nodes. There are 39,044 edges (sexual encounters) that are weighted with the rating between the buyer and seller. The weights are defined by bad (-1), neutral (0), and good (1). The edges contain attributes about the encounter, two of the attributes state whether or not the participants engaged in either anal or oral sex, with or without a condom. One attribute states whether the buyer and seller kissed or not. The last attribute states a number of days where the seller posted the listing until the buyer reviews the encounter.

After the filter of nodes with less than 2 edges and using a giant component, the network had 3,728 escorts and 5,420 buyers, which adds up 9,148 nodes. There are 31,878 edges.

4.3 Action Analytics

Escort as a profession is very dangerous and one bad encounter can lead to a sexually transmitted disease being exchanged. Since it is a legal profession in Brazil, the chances of STD outbreaks will be even greater. As an alternative path to the real world, we wanted to recommend an escort to a buyer, based on reviews and the buyer's preference. The first experiment will be using collaborative filtering and provide recommendations, which are defined by a few rules to determine the score. Some of the code snippets where provided by Toby Segaran [5] and Marcel Carciolo [6].

The scores are ranked from zero through ten, zero being the lowest, and ten being the highest. The rules are composed of the encounters rating value and use of protection during a sexual activity.

```
# Define final raiting
fgrade = 0.0
score = 0.0
if grade == '-1':
    fgrade = 1.0
    if anal == '1':
        score += 0.5
    elif anal == '-1':
        score -= 0.10
    if oral == '1':
        score += 0.5
    elif oral == '-1':
        score -= 0.10
    if kiss == '1':
        score += 0.5
    elif kiss == '-1':
        score -= 0.10
elif grade == '0':
    fgrade = 4.0
    if anal == '1':
        score += 0.75
    elif anal == '-1':
        score -= 0.20
    if oral == '1':
        score += 0.75
    elif oral == '-1':
        score -= 0.20
    if kiss == '1':
        score += 0.75
    elif kiss == '-1':
        score -= 0.20
elif grade == '1':
    fgrade = 7.0
    if anal == '1':
        score += 1
    elif anal == '-1':
        score -= 0.20
    if oral == '1':
        score += 1
    elif oral == '-1':
        score -= 0.20
    if kiss == '1':
        score += 1
    elif kiss == '-1':
        score -= 0.20
```

Figure 3. Rules that Define Final Score

The code snippet above showcases all of the rules. The first branches are defined on the grade that the buyer gave the seller. To start off from the top, 'if grade == -1' defines the fgrade as 1.0 because this is saying the encounter was a negative experience. The next two branches are 'grade == 0' and 'grade == 1', which give the fgrade a larger value because the grade given the encounter was either neutral or a good experience. Each main branch has six sub-branches that cater to the use of protection and user preference. If the activity included protection, then the score got incremented, but if the activity did not include protection, then the value got decremented. Both the fgrade and score values are added together to produce the final grade value.

'8200': {'3882': 9.0, '8001': 8.0}

Figure 4. Encounter Data Structure

The figure above shows an example how the data will be processed. From the left, the key`8200` is the node of the male buyer. Moving to the value object contains two encounters that share the same key and value. The keys `3882` and `8001` are the node id values of the sellers, and the values `9.0` and `8.0` are the final grade value.

Running some of the analysis code, the results will show the top recommendations. Using the male buyer node '370' the results show the top 5 female sellers.

[(9.000000000000002, '9291'),
(9.000000000000002, '6190'),
(9.000000000000002, '5224'),
(9.000000000000002, '4817'),
(9.000000000000002, '4615')]

Using the similarity distance parameter, the results varied due to using a different algorithm to determine the similarity. The left value is the score value, and the right value is the female seller node id.

[(9.000000000000002, '9064'),
(9.000000000000002, '8617'),
(9.000000000000002, '8132'),
(9.000000000000002, '5782'),
(9.000000000000002, '4418')]

Flipping the data structure, analysis can be ran to recommend a female seller to a male buyer. Using the female seller node '18' the results show the top 5 male buyers.

[(9.000000000000002, '9823'),
(9.000000000000002, '9136'),
(9.000000000000002, '9068'),
(9.000000000000002, '7261'),
(9.000000000000002, '4387')]

When using the similarity distance parameter, the top results showed different results.

[(9.000000000000004, '12363'),
(9.000000000000002, '9675'),
(9.000000000000002, '9584'),
(9.000000000000002, '8527'),
(9.000000000000002, '8492')]

Since the profession of an escort is legal in Brazil, the action analysis conducted in our study was able to gather some insights on using large datasets to recommend a male buyer a suitable female escort. In addition, our research was able to recommend a female seller a suitable male buyer.

4.4 Network Analytics

We used network analysis to simulate a viral outbreak within our data. The idea was borrowed from a paper written by Luis [7]. Several experiments were conducted, each introducing new parameters such as an additional probability that the virus will spread on an encounter, or change a variable and while also analyzing different viruses or sexually transmitted diseases. We started with HIV which has been found at a rate of 4.9% amongst sex workers in Brazil [2]. To simulate this, we used a random generator to initially "infect" 191 women, or 4.9% of the females in our dataset. The initial assumption being the rest of the dataset is "clean."

Our first experiment was simple in that it assumed a 100% success rate of the disease spreading, regardless of whether a condom was used, or the sexual act that was performed. Using a recursive algorithm that we created, we traversed through the graph to simulate the viral outbreak. For each "wave" or round the infected nodes would follow its own edges and "infect" its neighbor nodes. After 90% of the total population, the simulation would end. The result is shown below:

Table 3. Viral Outbreak Assuming No Contraceptive Usage

Wave	Total Infected	Newly Infected
0	191	0
1	1496	1305
2	4284	2788
3	8280	3396

Within three waves the over 90% of the graph was infected. This means that there are three degrees of separation between the initially infected group and the rest of the dataset. This is in line with the average shortest path being 5.18 nodes between any two nodes. With an average degree of ~7, the virus can exponentially spread throughout the graph. To simulate a real-world environment, we added an additional element of probability that the encounter will actually result in the virus spreading. The experiment was conducted at different intervals, 75%, 50%, 33%, 25%, 10%, and 1%, and the total number of waves and growth rate were calculated. The formula and results are below:

$$GrowthRate = \frac{(TotalInfected - InitiallyInfected)}{NumberOfWaves} * 0.01$$

Figure 5. Growth Rate Formula

Table 4. Viral Outbreak vs. Varying Condom Usage Rates

Probability (%)	Number of Waves to Infect 90%	Growth Rate
75	4	21.58%
50	5	17.05%
33	6	13.72
25	7	11.52%
10	15	5.36%
1	140	.57%

The result of this experiment was fruitful in determining an accurate probability that the virus would spread. This experiment was completed to account for a gap in the data provided, the usage of condoms. Since this data wasn't available, we could use probability to represent the safe sex that was occurring. We went with a probability of 33% in our subsequent experiments. Our experiment then analyzed having different initially infected individuals based off having the highest and lowest degree centrality, degree betweenness centrality, degree betweenness edge centrality, and degree closeness centrality. With these different starting nodes, we hoped to see how well the individuals

were connected to the network and the effect that had on the viral spread of the disease.

Table 5. Comparison of Centrality on Viral Outbreaks

Wave	High Centralities Total Infected	Lowest Centralities Total Infected
1	183	4
2	732	5
3	2911	9
4	4854	27
5	6472	87
6	7611	220
7	8260	422
8	N/A	1211
9	N/A	2780
10	N/A	5198
11	N/A	6865
12	N/A	7891
13	N/A	8458

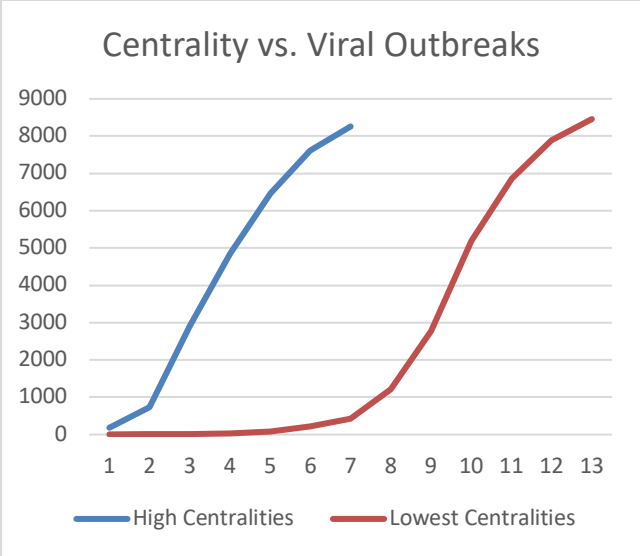

Figure 6. Comparison of Centrality on Viral Outbreaks

Interestingly enough the lower centralities starting nodes nearly took twice as many waves to spread the disease than the higher centrality starting nodes. It's not surprising to see that the growth rate mirrors this trend at 6.5% to 11.79% respectively. If we look at this as if it was viral marketing then it is quite simple to tell which starting nodes would be most susceptible to spreading the disease at the fastest rate. The working women with the most business clientele or men with the most amount of escort services purchased are the individuals who benefit from HIV and STD testing to help prevent the further spread of diseases.

Our last experiment analyzed different sexually transmitted diseases to see how fast each respective disease would spread among our dataset. From our previous experiment, we will use a probability of 33% as our underlying "spread rate" parameter. The viruses/diseases we will track include HIV, HPV, Chlamydia, and any STD in general. Then initial prevalence among female escorts is: 4.9%, 67.7%, 20.5%, 71.6% respectively [3].

Table 6. Effect of Different Diseases on Growth Rate

Virus/Disease	Initially Infected	Number of Waves to 90% Infection	Growth Rate
HIV	191	6	13.72%
HPV	2502	3	19.69%
Chlamydia	770	5	15.46%
Any STD	2678	3	19.37%

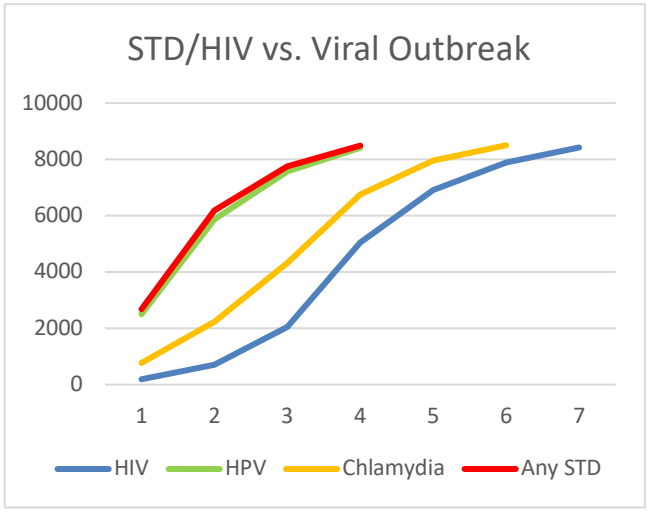

Figure 7. Comparison of STD/HIV on Viral Outbreaks

5. Conclusion

In conclusion, the sex trade profession in Brazil is legal and here to stay. The lessons learned from our experiments can not only help this industry become safer for both the buyers and suppliers, but also can also provide insight to the viral spread of HIV and STDs which have become rampant in certain impoverished areas of Brazil. Using action analytics, we were able to create a recommendation system by giving both men and women in our dataset a score. For escorts, this meant determining which male would be safest to solicit to. For the buyers, it factored in both cleanliness and the rating the females received. Using network analytics, we were able to simulate a viral outbreak to see which factors had the greatest effect on the spread of a disease. We determined that usage of a contraceptive would greatest decrease the spread, as it drastically decreases the probability of infection on an encounter. We also determined that diseases that have already been widespread throughout the community have a greater growth rate. If the government were to put in place programs that would provide testing of these well-known epidemics, safer sex could be practiced among these individuals. We hope our results

can be used to control this widespread epidemic, and make this industry both safer for all parties involved in the transaction.

For a future study, having text review comments in addition to the network data would improve the results of the recommendations. Text data would be analyzed through sentiment analysis to determine the encounters semantic orientation and would be added to the rules to define the final review score. Topic modeling could be determined to identify the top topics to cater a buyer or seller's preference. Analyzing the text data could provide some interesting insights on what other sentiments or topics are gained when highlighting keywords that full under 'use of protection during an encounter'.

In the future, if our dataset had a few more parameters we could fine tune our experiment to more accurately reflect the real world. The first data point in our experiment would benefit from having a timestamp of when the sexual escort was purchased. This would add another dimension to traversing our graph, the timestamps could determine the number of "waves" before the virus spread throughout the population. The arbitrary concept of waves, in fact, could be replaced by a more accurate 'time' or 'duration' in which the viral outbreak spreads. Another crucial piece of information the experiment would benefit from is having data on proper condom usage. The usage of condoms would determine the probability of a sexually transmitted disease actually spreading during an encounter. Using a condom drastically lowers the probability that the disease would spread. Lastly, we used data collected from studies to determine the initial number of affected individuals, however, if we had data from STD and HIV tests, we could more accurately determine our initially infected.

6. REFERENCES

[1] Central Intelligence Agency. 2017. *Country Comparison :: HIV/AIDS – Deaths.* (April 2017). Retrieved April 22, 2017 from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2157rank.html#br

[2] Advert. 2017. *HIV and AIDS in Brazil.* (April 2017). Retrieved April 22, 2017 from https://www.avert.org/professionals/hiv-around-world/latin-america/brazil

[3] Christina Maria Gracia de Lima Parada. 2011. *HIV and AIDS in Brazil.* (June 2011). Retrieved April 22, 2017 from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-11692011000300007

[4] Jérôme Kunegis. 2016. *KONECT - The Koblenz Network Collection.* (October 2016). Retrieved February 18, 2017 from http://konect.uni-koblenz.de/networks/escorts

[5] Marcel Caraciolo. 2009. *Collaborative Filtering: Implementation with Python*! (November 2009). Retrieved April 22, 2017 from https://aimotion.blogspot.com/2009/11/collaborative-filtering-implementation.html

[6] Toby Segaran. *Programming Collective Intelligence: building Smart Web 2.0 applications*. O'Reilly Media, 2007.

[7] Luis E. C. Rocha, Fredrik Liljeros, Petter Holme. 2011. *Simulated Epidemics in an an Empirical Spatiotemporal Network of 50,185 Sexual contracts*. (March 2011). Retrieved February 18, 2017 from http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001109