Arrest Charges' (Non)-Independence in Black and White Men

An Extension and Application of SimRank to include Network Structure when Computing Black and White Men's Arrest Charge Similarities

Dustin Fink (dsfink) and Katharina Roesler (roesler)

12/08/2015

1 Abstract

When people are arrested, they are frequently charged with several offenses at once, yet most criminological research focuses on charges individually, as though they were independent of one another. This project examines how charges relate to each other within a charge-arrestee network, and how charge similarity differs across race.

We extend the SimRank algorithm (Jeh and Widom 2002) to efficiently examine the similarity of charge pairs in a large dataset of men arrested in 2013. We compare computed SimRank scores of charge pairs within black and white arrestees and evaluate how these measures differ from the more traditional measure of similarity of Chi-squared residuals, which does not account for network structure. Our procedure identified certain charges involving drugs, vandalism, theft, and assault to have a higher similarity in terms of the network than would be assumed based on their simple co-occurrences, and found that these measures were similar across black and white men.

2 Introduction

Most criminological research examines arrest charges independently. That is, studies tend to examine individuals arrested for a certain type of charge, ignoring other charges in the same arrest. However, individuals are often charged with multiple offenses during a single arrest. It is of interest which charges co-occur frequently and whether co-occurrence patterns vary by arrestee race.

For instance, certain patterns of co-occurrence and their interaction with arrestee race may indicate biased policing. That is, if stereotypically black charges are related to a more diverse set of other charges for black rather than white arrestees, it may be that Black individuals are under heightened police scrutiny for these offenses. In addition, if charges for which officers have relatively greater discretion to arrest individuals are related to a more diverse set of charges for black than for white arrestees, then officers may be over-arresting black men or under-arresting white men. The former was certainly the case in Ferguson, Missouri, where officers disproportionately charged black residents with a variety of misdemeanors, such as "manner of walking" and "failure to comply" (U.S. DOJ 2015, 67).

One can obtain a baseline measure of similarity by examining how often two charges are present in a single arrest. However, we can extend this concept of similarity by considering

charges more similar when people arrested for both are more similar along other dimensions as well. This measure of similarity is given by the SimRank algorithm. We begin with a discussion of SimRank, extend it to run efficiently on our data, and give a procedure for determining which SimRank scores are significant due to network structure as opposed to the frequency of charges or simple co-occurrences.

3 SimRank: A Measure of Structural-Context Similarity

SimRank (Jeh and Widom 2002) provides an algorithm for assigning similarity between pairs of vertices in any given network based on their structural contexts. Two vertices are considered similar if they are connected to nodes which themselves are similar. This recursive definition of similarity improves upon previously used measures of similarity for networks (which only compared the shared number of neighbors between nodes) by allowing similarity to propagate through the network.

SimRank strengthens measures of similarity on the assumption that similarity is a quality that can propagate through a network. If this assumption makes sense for the modeled data, then SimRank is an effective and meaningful measure of similarity based on complex contextual structures. SimRank is quadratic in the size of the graph (and for dense graphs, cubic), so the run-time is not ideal. Run-time improvements can be achieved by pruning and advanced disk-writing techniques, but these either reduce the quality of the results or require more complex engineering.

4 Problem Statement

We develop and propose a strategy for identifying which offenses have higher similarity due to their network characteristics. We hypothesize that black and white male arrestees charges have different network structures. More precisely, we hypothesize that within each racial group, different charges are similar to one another, due to different patterns of criminal behavior and policing for each racial group.

5 Strategy

1. Calculate SimRank scores between charges, made feasible by our Compressed SimRank algorithm.
2. Standardize the observed SimRank scores by subtracting the SimRank scores from a graph with the same single charge frequencies but randomized co-occurrences.
3. Compare these standardized SimRank scores with the baseline similarity of residuals of a chi-squared test done on co-occurrences to see which charges are more similar when considering network context.
4. Repeat the above restricted to black and white populations, and contrast the findings.

6 Data

The Federal Bureau of Investigation collects data from law enforcement agencies on a monthly basis, producing an annual file titled the National Incident Based Reporting System (NIBRS; National Archive of Criminal Justice Data 2013). In this project we examine NIBRS arrestee data from 2013, which contain records for 2,361,308 arrested men. Because NIBRS relies on agencies to voluntarily report their statistics, there may be considerable selection bias in these data.

We merge this arrestee-level file with data on law enforcement agencies from the U.S. Department of Justices Law Enforcement Management and Administrative Statistics (LEMAS) survey (U.S. D.O.J. 2013). This dataset allows us to compare arrestees and their charges at the individual level while also incorporating information about their social context.

Of the 2.4 million arrested men included in NIBRS data, we exclude records for which the charges are unknown. This results in a dataset of 1,124,876 arrested men (white: 744,889; black: 343,706).

7 Models

We take advantage of three models: the arrestee, the profile model, and the random-profile model.

7.1 Arrestee Model

We model our data as an undirected, bipartite graph $G(O, H, E)$ where

- O = Node set of all arrestees
- H = Node set of all charges
- E = Edge set linking each arrestees with the corresponding attributes and charges.

Each offender is charged with 1-3 charges.

7.2 Profile Model

In the profile model, we profile arrestees by their set of charges, and collapse all arrestees with the same profile into one profile node.

We represent our compressed network an undirected, bipartite graph $G(P, H, E)$ where

- P = Node set, where each node represents a profile p with $|p|$ offenders.
- H = Node set of all charges
- E = Edge set linking each profiles with the corresponding attributes and charges.

The number of observed profiles were 1,995 for the total data, 1,775 for the white subset, and 1,312 for the black subset, out of 16,261 profiles possible profiles.

7.3 Random-Profile Model

We generate a random-profile graph that has the same counts of arrestees for each charge, but randomized the arrestees profiles. We generated this graph by iteratively selecting 1-3 charges and adding an arrestee to the corresponding profile while decreasing the corresponding charge counts. The algorithm ran until all charge counts were 0. This produced a graph with a similar number of profiles (total: 2,561 total; white: 2,320; black: 1,963), but a different distribution of charges.

8 Methods

8.1 SimRank

Following the original methods for finding SimRank for a bipartite graph with $C = C_1 = C_2 = 0.9$, for all $a, b \in V, V = H, O$, we start by setting

$$s_0(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}$$

We then iterate using the update equation:

$$s_k(a,b) = \begin{cases} 1, & \text{if } a = b \\ \frac{C}{|I(a)||I(b)|} \sum_{\substack{c \in I(a) \\ d \in I(b)}} s_{k-1}(c,d), & \text{if } a \neq b \end{cases}$$

until convergence of a maximum number of iterations. We say the algorithm converges when all SimRank scores are within ϵ of the previous iteration.

The runtime of this naive approach is is $O(K(|O|^2 d_O + |H|^2 d_H))$, where K is the number of iterations, and d_V is the average value of $|I(a)||I(b)|$ for $a, b \in V$.

8.2 Compressed SimRank

SimRank is prohibitively large given the number of offender in our dataset. The original SimRank paper suggests pruning as a way to increase computation, but this can decrease the quality of the results. Instead, we can use our profile model along with what we describe here as Compressed SimRank to achieve the correct SimRank scores in a feasible amount of time.

Compressed SimRank relies on the fact that any pair of offenders from the same pair of profiles will have the same SimRank score

$$s(a,b) = s(a',b') \text{ for } a,b \in p, a',b' \in q, a \neq b \text{ and } a' \neq b'$$

since they have the same neighbor sets.

We can thus compute SimRank scores on the profile model instead of the offender model. We define the SimRank between two profiles to be the SimRank of two offenders taken from those profiles. The SimRank of a profile with itself is the SimRank of two different offenders from that profile (The SimRank of a profile that contains only one offender with itself does not affect the equations, and thus is calculated as if there were more than one offender).

Using our compressed network, we change the initial equation for profiles to be:

$$s_0(p,q) = 0$$

and the initial equation for charges to be the same:

$$s_0(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}$$

The update equation for profiles changes to:

$$s_k(p,q) = \frac{C}{|I(p)||I(q)|} \sum_{\substack{c \in I(p) \\ d \in I(q)}} s_{k-1}(c,d)$$

and the update equation for charges changes to

$$s_k(a,b) \quad = \quad \frac{C}{\sum_{\substack{p \in I(a) \\ q \in I(b)}} |p||q|} \left[\sum_{\substack{p \in I(a) \\ q \in I(b) \\ p \neq q}} |p||q| s_{k-1}(p,q) + \sum_{\substack{p \in I(a) \\ q \in I(b) \\ p = q}} (|p|^2 - |p|) s_{k-1}(p,q) + |p| \right]$$

if $a \neq b$ and $s_k(a,b) = 1$ if $a = b$.

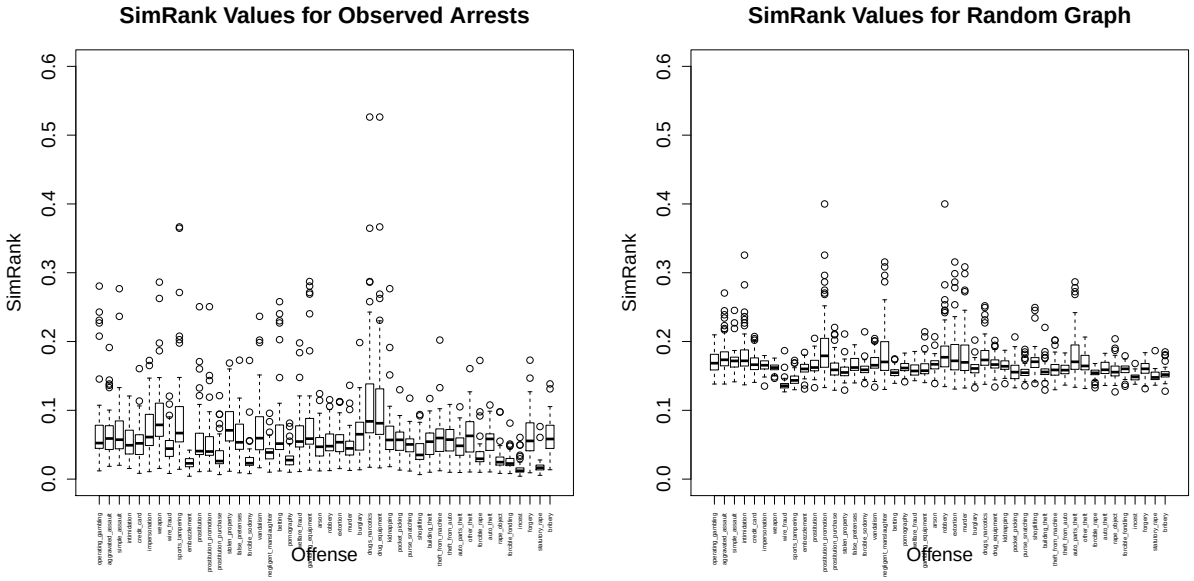
The resulting scores for pairs of charges is the same as SimRank. We can also recover the SimRank of a pair of two different offenders from the SimRank of their corresponding profiles. The derivation of these equations is given in Appendix A.

The resulting runtime is $O(K(|P|^2 d_P + |C|^2 d_C)) \sim O(K(|P|^2))$. Since the number of profiles is about 2/1000ths the number of offenders, this runs in much faster time, and finished in 1-2 hours.

Compressed SimRank was run on the original data, the black subset, and the white subset, as well as a random-profile graph based on the single charge frequencies of the original data, the black subset, and the white subset.

9 Evaluation

We "standardize" computed SimRank scores for the observed data by subtracting the SimRank scores for the random-profile graphs. The figures below show the distributions of SimRank scores for each offense for the observed data on the left and the random network's data on the right. SimRank scores for the observed data have greater variance than do those for the random network, indicating that charges are more related to one another than one would expect by chance. The random SimRank scores have a higher mean because the randomization resulted in more profiles with greater sizes, and thus more simulated offenders, but this difference does not affect our evaluation.

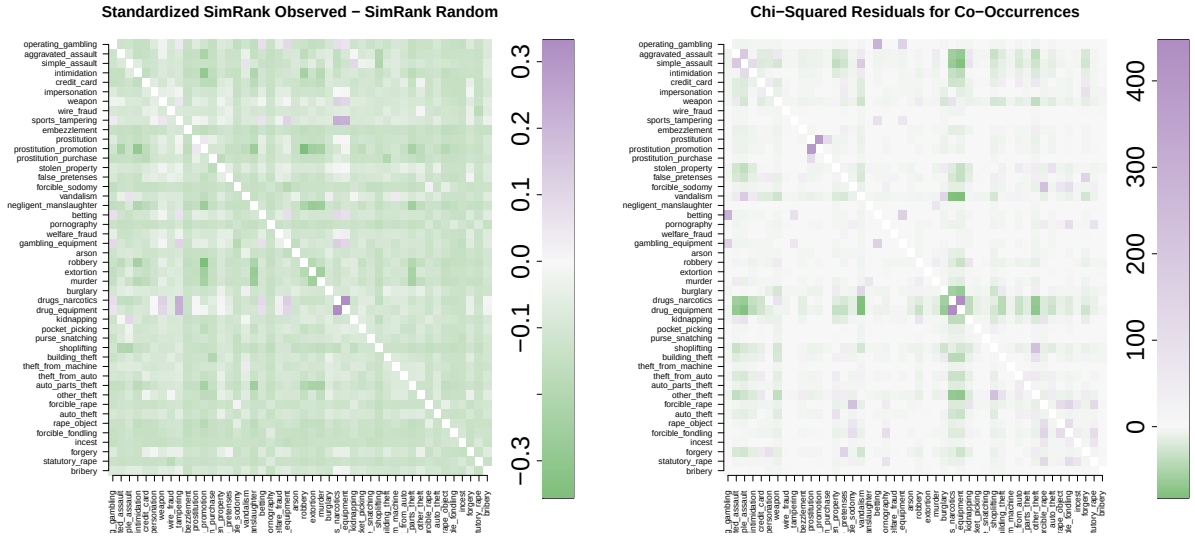


In order to evaluate the general accuracy of the computed SimRank measures, we compare charge-pairs' similarity scores and Chi-squared residuals in the raw, co-occurrence data. For instance, to assess whether White men's relatively high standardized similarity score (0.121) for weapon and drug charges is accurate, we calculate a Chi-squared table for the two charges. Given their respective frequencies, their expected co-occurrence is 3,835, while their actual co-occurrence is 7,915. This results in a Chi-squared residual of 65.9 for co-occurrence and a general p-value of 0.0000, indicating that the two charges co-occur much more frequently than expected by chance.

Black men have a relatively lower standardized SimRank score for weapon and drug charges (0.080), although this is still well above black men's average standardized SimRank of all charge pairs (-0.149). For black arrested men, the expected co-occurrence of weapon and drug charges is 4,155, while the observed co-occurrence is 7,919. This results in a Chi-squared residual of 58.4 and a general p-value of 0.0000, indicating that the two charges are related, but slightly less so than they are for white men.

In order to compare all computed SimRank values to observed co-occurrences simultaneously, we plotted matrices of charge-pairs' standardized SimRank scores and Chi-squared

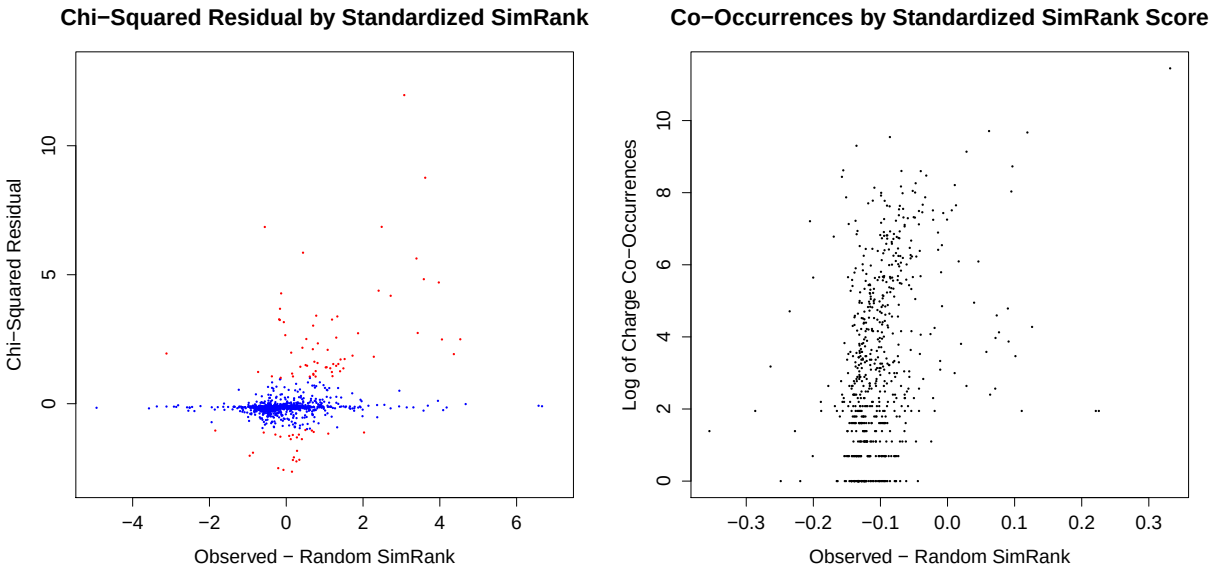
residuals. The figure below on the left below shows our standardized SimRank scores: the difference in SimRank scores between the observed and random networks. It seems that drugs/narcotics and drugs equipment are highly related to one another and many other charges, while most other charge pairs are less related than the random network predicts.



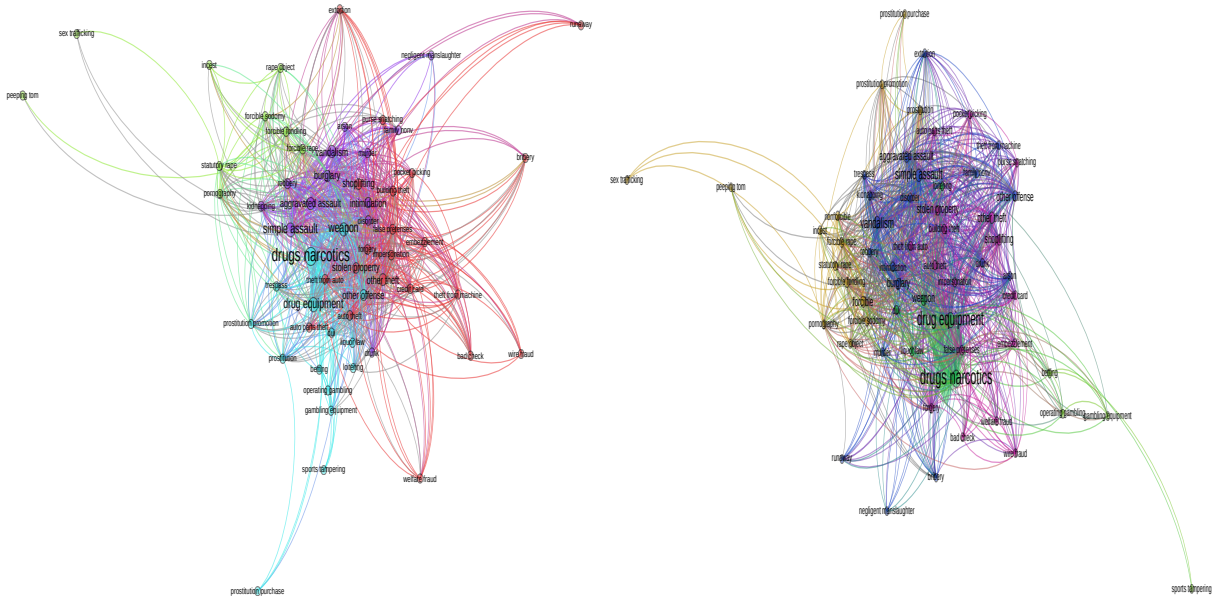
A similar pattern emerges with the Chi-squared residuals, shown in the figure on the right. Drugs and narcotics charges are highly related to one another, while most other charges are largely unrelated to one another.

The figure below on the left compares charge pairs' standardized SimRank scores and Chi-squared residuals. Blue dots represent charge pairs whose Chi-squared residual is between -1 and 1, while red dots represent charge pairs with relatively larger Chi-squared residuals (in absolute terms). It seems that SimRank closely matches Chi-squared residuals for charges that co-occur much more or less often than expected, but does not do so for charge pairs that are relatively unrelated. This makes sense, as network structure impacts SimRank scores more strongly when charges are unrelated to one another.

SimRank is thus somewhat independent of charge co-occurrence. The figure below on the right confirms this independence, as charge pairs' standardized SimRank scores are only moderately related to the log of their co-occurrences.



In addition to performing these analyses, we examine our data visually to understand how black and white arrestees' charges relate to one another. We create a charge-charge network with edges weighted by charges' co-occurrence. That is, for each individual arrested for charge 1 and charge 2, the edge between charge 1 and 2 increases by 1 in weight. Please see the figures below, which represent the charge-charge networks for black and white arrested men, with Black men on the left.



Communities are detected by modularity and colored accordingly, and charges with greater

weighted PageRank have larger labels and are more central to the network. Both black and white arrestees' networks have similar communities and charges high in centrality, mirroring their similar SimRank scores.

9.1 Final Results and Discussion

We compare the distributions of standardized SimRank scores and Chi-squared residuals and find that the following charge pairs are two standard deviations above the mean (are "significantly" higher in SimRank):

- drugs equipment, with each of vandalism, simple assault, aggravated assault, burglary, other theft
- drugs narcotics, with each of vandalism, other theft, simple assault

In addition, the following charge-pairs are six standard deviations below the mean (are relatively lower in SimRank):

- drugs/narcotics with drugs equipment
- prostitution with prostitution promotion
- gambling operation with betting
- forcible sodomy with forcible rape
- shoplifting with other theft

The implication here is that our procedure identifies an important contribution of the network structure to the similarity of vandalism, theft, assault, and drug crimes. The pairs below the mean difference have relatively large standardized SimRank scores due to their high count and/or simple co-occurrence. We present pairs six deviations below the mean due to space constraints, but the tails are similar and the large majority of pairs (96%) are within 2 standard deviations.

We calculate the same distributions for subsets of black and white arrestees separately and find similar charge-pairs to be relatively high and low in SimRank as compared to Chi-squared residuals. Thus we find no support for our hypothesis that charges' similarities vary by arrestee race.

10 Conclusion

We estimate a measure of charge similarity that is related to but also divergent from charges' co-occurrences. We compute Compressed SimRank to enhance classical SimRank, and standardize these scores against those of a comparable random graph. High standardized SimRank scores relative to co-occurrence indicate that a pair of charges share similar contexts. We compare these measures for black and white arrestees and find that black and white arrestees' charges relate to one another in similar ways.

11 References

Jeh, Glen, and Jennifer Widom. "SimRank: a measure of structural-context similarity." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

National Archive of Criminal Justice Data. *National Incident-Based Reporting System, 2013: Extract Files.* ICPSR36121-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-08-04. http://doi.org/10.3886/ICPSR36121.v1

United States Department of Justice (DOJ). Office of Justice Programs. Bureau of Justice Statistics. *Law Enforcement Management and Administrative Statistics (LEMAS), 2013, ICPSR 36164.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor], 2015. doi.org/10.3886/ICPSR36164.v1.

United States Department of Justice (DOJ). 2015. "Investigation of the Ferguson Police Department." Technical report, United States Department of Justice, Civil Rights Division.

12 Teammate Contributions

Dustin: Models and Algorithms, Compressed SimRank, Random-profile generation, Final Results and Discussion

Katharina: Data cleaning, Data visualization, Evaluations and evaluation visualization

The following pages are included for completeness' sake and not intended to be graded.

Appendix A - Derivation of Compressed SimRank

Let $G(O, H, E)$ be an undirected bipartite graph, and $G(P, H, E)$ be the compressed profile version of the graph.

Let SimRank for the original graph to be defined by the equations

$$s_0(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}$$

and

$$s_k(a,b) = \begin{cases} 1, & \text{if } a = b \\ \frac{C}{|I(a)||I(b)|} \sum_{\substack{c \in I(a) \\ d \in I(b)}} s_{k-1}(c,d), & \text{if } a \neq b \end{cases}$$

We take it as fact that

$$s(a,b) = s(a',b') \text{ for } a,b \in p, a',b' \in q, a \neq b \text{ and } a' \neq b'$$

for profiles $p, q \in P$. Noting this, we can then define the SimRank between to profiles to be

$$s(p,q) = s(a,b)$$

for any $a \in p, b \in q, a \neq b$. This definition is ill-defined if $p = q$ and $|p| = 1$, but this will not end up mattering in our derivation. For the purposes of the rest of the derivation, we let a, b be taken from p, q with $a \neq b$.

Given this definition, our initial values for $s(p,q)$ are 0, since

$$s_0(p,q) = s_0(a,b) = 0$$

Our update rule for profiles is

$$s_k(p,q) = s_k(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{\substack{c \in I(a) \\ d \in I(b)}} s_{k-1}(c,d)$$

Let $\hat{I}$ denote the neighbor set in our compressed graph. Because of our compression, we have that $I(a) = \hat{I}(p)$ and $I(b) = \hat{I}(q)$, so we can re-write this as

$$s_k(p,q) = \frac{C}{|\hat{I}(p)||\hat{I}(q)|} \sum_{\substack{c \in \hat{I}(p) \\ d \in \hat{I}(q)}} s_{k-1}(c,d)$$

It remains to derive the update equation for charges in Compressed SimRank in terms of the SimRank scores of profiles from the original SimRank update equation.

For a pair of charges c, d, if $c = d$, we still have $s_k(c, d) = 1$. Otherwise, we have

$$s_k(c,d) = \frac{C}{|I(c)||I(d)|} \sum_{\substack{a \in I(c) \\ b \in I(d)}} s_{k-1}(a,b)$$

We first note that the neighbor set of a charge c is the union of all offenders of all profiles linked to c in the compressed graph:

$$I(c) = \cup_{p \in \hat{I}(c)} \cup_{a \in p} a$$

We thus get

$$\begin{aligned} |I(c)| &= |\cup_{p \in \hat{I}(c)} \cup_{a \in p} a| \\ &= \sum_{p \in \hat{I}(c)} |\cup_{a \in p} a| \\ &= \sum_{p \in \hat{I}(c)} |p| \end{aligned}$$

and similarly for d with q. The first equality works since the profiles partition the set of offenders.

We next consider the sum $\sum_{\substack{c \in \hat{I}(p) \\ d \in \hat{I}(q)}} s_{k-1}(c,d)$. We can regroup the offenders by profile to get

$$\sum_{\substack{a \in I(c) \\ b \in I(d)}} s_{k-1}(a,b) = \left[\sum_{\substack{p \in \hat{I}(c) \\ q \in \hat{I}(d) \\ p \neq q}} \sum_{\substack{a \in p \\ b \in q}} s_{k-1}(a,b) + \sum_{\substack{p \in \hat{I}(c) \\ q \in \hat{I}(d) \\ p = q}} \left[\sum_{\substack{a \in p \\ b \in q \\ a \neq b}} s_{k-1}(a,b) + \sum_{\substack{a \in p \\ b \in q \\ a = b}} s_{k-1}(a,b) \right] \right]$$

Since all SimRank scores are the same for a pair of offenders from the same pair of profiles, we can replace the inner sums with products of the SimRank of the profiles and the profile sizes.

$$\sum_{\substack{a\in I(c)\\ b\in I(d)}} s_{k-1}(a,b) = \left[\sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p\neq q}}\sum_{\substack{a\in p\\ b\in q}} s_{k-1}(a,b) + \sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p=q}}\left[\sum_{\substack{a\in p\\ b\in q\\ a\neq b}} s_{k-1}(a,b) + \sum_{\substack{a\in p\\ b\in q\\ a=b}} s_{k-1}(a,b)\right]\right]$$

$$= \left[\sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p\neq q}} |p||q| s_{k-1}(p,q) + \sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p=q}} \left[(|p|^2-|p|)s_{k-1}(p,q) + |p|^2(1)\right]\right]$$

We separate the terms like this to account for the case when $p = q$ and $a = b$, since the value of $s(a, b)$ in this case is 1 and not $s(p, q)$. Note that if $p = q$ and $|p| = 1$, $|p|^2 - |p| = 0$ and so the undefined value of $s(p, q)$ in this case does not matter.

Putting it all together, we get

$$s_k(c,d) = \frac{C}{|I(c)||I(d)|}\sum_{\substack{a\in I(c)\\ b\in I(d)}} s_{k-1}(a,b)$$

$$= \frac{C}{\sum_{p\in\hat{I}(c)}|p|\sum_{p\in\hat{I}(d)}|q|}\left[\sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p\neq q}} |p||q| s_{k-1}(p,q) + \sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p=q}} \left[(|p|^2-|p|)s_{k-1}(p,q) + |p|^2(1)\right]\right]$$

$$= \frac{C}{\sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)}}|p||q|}\left[\sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p\neq q}} |p||q| s_{k-1}(p,q) + \sum_{\substack{p\in\hat{I}(c)\\ q\in\hat{I}(d)\\ p=q}} \left[(|p|^2-|p|)s_{k-1}(p,q) + |p|^2(1)\right]\right]$$