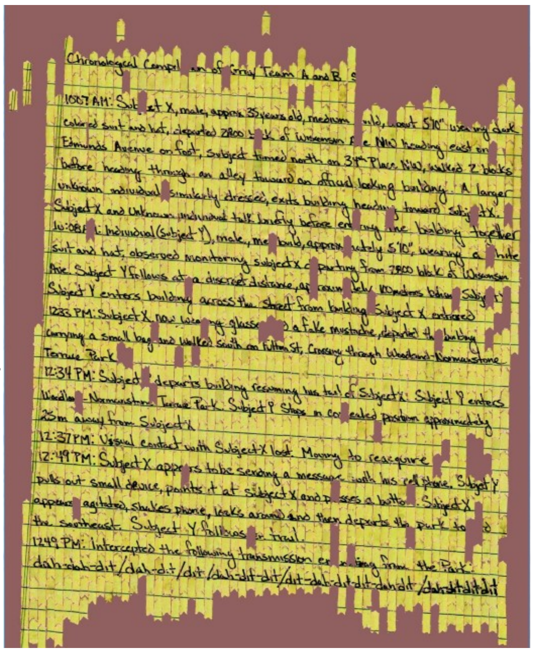
# Reconstructing shredded documents

## Problem:

Ever since the paper shredder was invented, people have worked on sticking the pieces back together again. Techniques permitting the purely electronic storage and transmittal of sensitive documents have been developed but, because of convenience or for legal reasons, many such documents are still printed and eventually shredded.

The recent surge of interest in the area of reconstructing shredded documents has shown that many people overestimate the security offered by the shredder. One example is the 2011 DARPA shredding challenge, which offered a $50,000 prize for the first team to successfully solve a series of  puzzles printed on 5 shredded documents. When a winner was declared after 32 days, the performance was generally met with amazement from the general public, especially since the final puzzle in the contest (pictured left) is on par with documents shredded by some of the most expensive devices available today.


Chronological Compil an of Gray Team A and B s

10:07 AM: Sub ct X, male, approx 35 years old, medium ild, out 5'10" wearing dark colored suit and hat, departed 2800 k of Wisconsin e NW heading east on Edmunds Avenue on foot, Subject turned north on 34th Place NW, walked 2 blocks before heading through an alley toward an official looking building. A larger unknown individual similarly dressed, exits building heading toward Subj tX. Subject X and Unknown individual talk briefly before en ring ne building together
10:08 : Individual (Subject Y), male, me build, approx ately 5'10", wearing a hite suit and hat, observed monitoring subject X  parting from 2800 block of Wisconsin Ave. Subject Y follows at a discreet distance, ap roxim tely 10 meters behin Subje Subject Y enters building across the street from building Subject X entered
12:33 PM: Subject X, now wea ng glasse  a fake mustache, departed th building carrying a small bag and walked south on Fulton St, Crossing through Woodland-Normanstone Terrace Park
12:34 PM: Subject departs building resuming his tail of Subject X. Subject Y enters Woodl -Normanstone Terrace Park. Subject Y Stops in co ealed position approximately 25m away from Subject X
12:37 PM: Visual contact with Subject X lost. Moving to reacquire
12:49 PM: Subject X app rs to be sending a message with his cell phone. Subject Y pulls out small device, points it at Subject X and p sses a butto . Subject X appears agitated, shakes phone, looks around and then departs the park to d the southeast. Subject Y follows in trail.
12:49 PM: Intercepted the following transmission em ting from the Park:
dah-dah-dit/dah-dit/dit/dah-dit-dit/dit-dah-dit-dit-dah-dit/dah-dit-dit-dit

**DARPA Challenge, reconstruction of part of puzzle 5**

## Applications:

The major application of unshredding is, of course, in forensics and investigative sciences.

The DARPA Challenge itself was motivated by the difficulty troops encounter in war zones while trying to make sense of the remnants of destroyed documents.

One historic case where unshredding technology would have been of direct use is the Iranian Revolution in which, after the takeover of the U.S. embassy in Tehran, Iranian authorities hired carpet weavers to reconstructed the documents by hand

Probably the most ambitious shredded document reconstruction attempt is the current effort to recover the shredded archives of the East German secret police. There are a total of 16,000 bags of shredded documents and, so far, it took three dozen people six years to reconstruct 300 of them. Without further computational assistance it would take 11520 man-years to finish this project.

There are also tangential applications in fields like archaeology, where attempts are made to put together pieces of artefacts. The major difference between the two problems is that in this latter case a lot more information can be inferred from the shapes of the pieces themselves. This is significantly different from shredded document recovery where all the information is in the content of the shreds rather than their shape.

## Approach in literature:

Almost universally in literature, the solution to the unshredding problem has been formulated with the help of two functions.

First, there is a cost function which looks at 2 shreds and assigns a number to them proportional to how good a match they are. If the cost function considered the 2 shreds a perfect match, the cost will be 0, if the match is impossible it will be infinite.

Second, there is a search function which, given all the costs over pairs of edges, tries to place the shreds in such a way that the sum of the costs of all adjacent pieces is minimal. This permutation is considered the most likely solution.

Significant effort has been made towards finding a good search function.  In [1] it is shown that the problem is NP hard by reduction to the Travelling Salesman Problem. In [2] an exact solution is attempted via Integer Linear Programming which yields very good results but is  intractable for any number of strips larger than 150.

Furthermore, several attempts ([3],[4] )have been made at applying evolutionary algorithms to the search problem.

In contrast, relatively little progress has been made in developing the cost function, most papers settling on a very simple formulation which does a weighted difference of the adjacent pixels on either side of the proposed join.

One of the few refinements proposed for the cost function is the one in [5], where the authors try to solve the problem of white on white joins which give a perfect match for the previous, naive, cost measure but in reality do not give a lot of information.

## My approach – general idea:

Most modern reconstruction approaches try to blend the computer's computational power with the human's knack for solving tough situations.

This means that most of the previously discussed methods optionally include a human in the loop, though the extent of the work the human is expected to perform varies greatly. The approaches attempted in the DARPA challenge, for instance, ranged from crowd sourcing the task on mechanical turk to mostly automated approaches that only relied on a human to flag mistakes made by the algorithm.

As opposed to this, I am looking at purely automatic methods and am trying to undertake a more principled approach to the problem as to be able to look at what the fundamental security of shredders is. This means I am looking at both some technical and some practical details that have been largely ignored so far.

On the practical side, most methods simplify some of the noise and complexities of the problem by assuming the shredding was done perfectly, without any torn or ragged edges, that the shreds are in good condition with no excessive bending or additional hand tearing and that we know a-priori exactly which pieces form a certain document. The removal of some of these assumptions can make the problem significantly harder.

On the theoretical side, the methods attempted so far either fall in the complete but intractable side (like the Integer Linear Programming approach in [2] and the Travelling Salesman reduction in [1]) or in the heuristic/local search side(as the genetic algorithms in [3] and [4]). More principled heuristic approaches are possible however which I believe will yield better results and will be more easily adapted to different document types.

## My approach - details:

I noticed that in [5], by adopting a slightly more sophisticated heuristic for the cost function, the authors have managed to surpass the performance of the evolutionary algorithm presented in [3] even though their actual search function is a simple greedy method. This shows there is much to be gained by further improvements to the cost function. However, there is no reason to limit ourselves to simple heuristics when we can apply well known machine learning approaches instead.

One approach I have been investigating is using the given shreds to identify similar documents from a database and then applying neural networks over these similar documents in order to train a cost function that will perform well on the original shreds. Another approach I am attempting is the use of Bayesian statistics to model the probability of two shreds matching given the distribution of edges encountered so far.

Both of these have the ability to simulate the simpler heuristics previously employed while also being ,pre heavily dependant on the particular shreds we have and thus more adaptable to varying datasets.

I am also looking at different search functions. For instance, i have implemented a heuristic that functions analogous to Kruskal's algorithm. Namely at first it considers each shred as a separate forest, it then picks the best match between two forests and merges them. It repeats this until there is only one forest left. The only difficulty is in estimating the cost of merges, because the positioning of the items in an 2D plane is important here whereas it is irrelevant for the classic Kruskal algorithm. For instance a merge which would results in two pieces being superimposed must have an infinite cost.

This method isn't present in literature and performs surprisingly well. Specifically, on instances over 10x10 it gets results comparable to those of the genetic algorithm in [3] at a fraction of the computational cost.

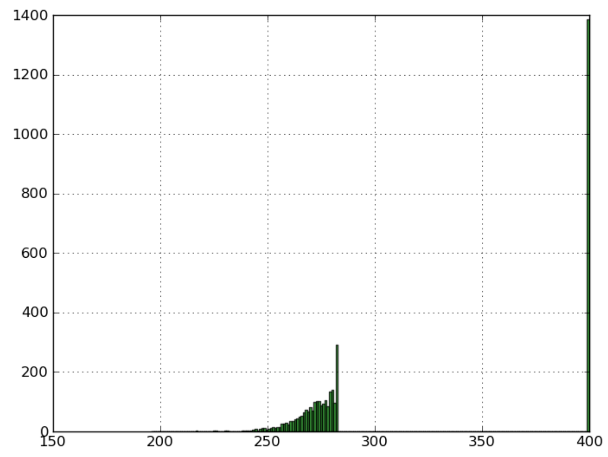
Additionally, this method could also be further improved by looking at the Bayesian probabilities of matches and adjusting the scores based on the fact that two odd shaped forests that happen to match together perfectly should give us a higher confidence of a correct match than two squares matching.

## Preliminary Results:

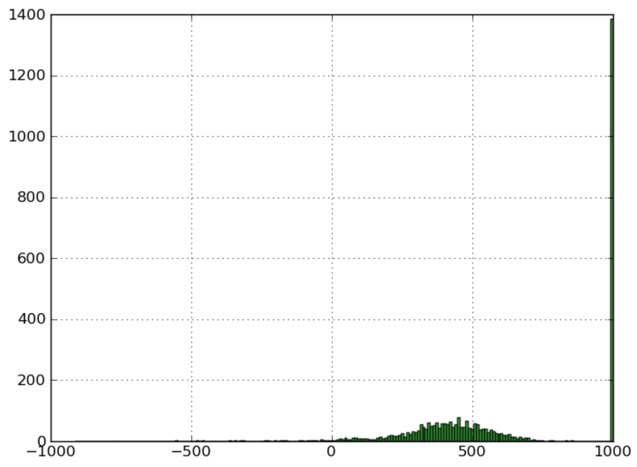
I have implemented several heuristics for both the cost and the search functions so far, including the above “kruskal” search function. As mentioned the novel “kruskal” search together with the cost heuristic used in [5] already perform at a level on par to that in papers [1] and [3], while being however surpassed by the results of papers [4] and [5].

I have also looked in some detail at the properties of several heuristics used in literature.

For instance, a closer investigation of the cost heuristics presented in [5] (graphs below) shows the advantage of smoothing the results by a Gaussian. The Gaussian is able to better separate the data such that fewer true positives and false positives fall in the same bucket. When applying Bayesian statistics to the cost function I would expect to get a similar distribution without having to explicitly use a smoothing function like a Gaussian.



**Distribution of scores from a bit to bit cost heuristic**



**Distribution of scores from a smoothed cost heuristic**

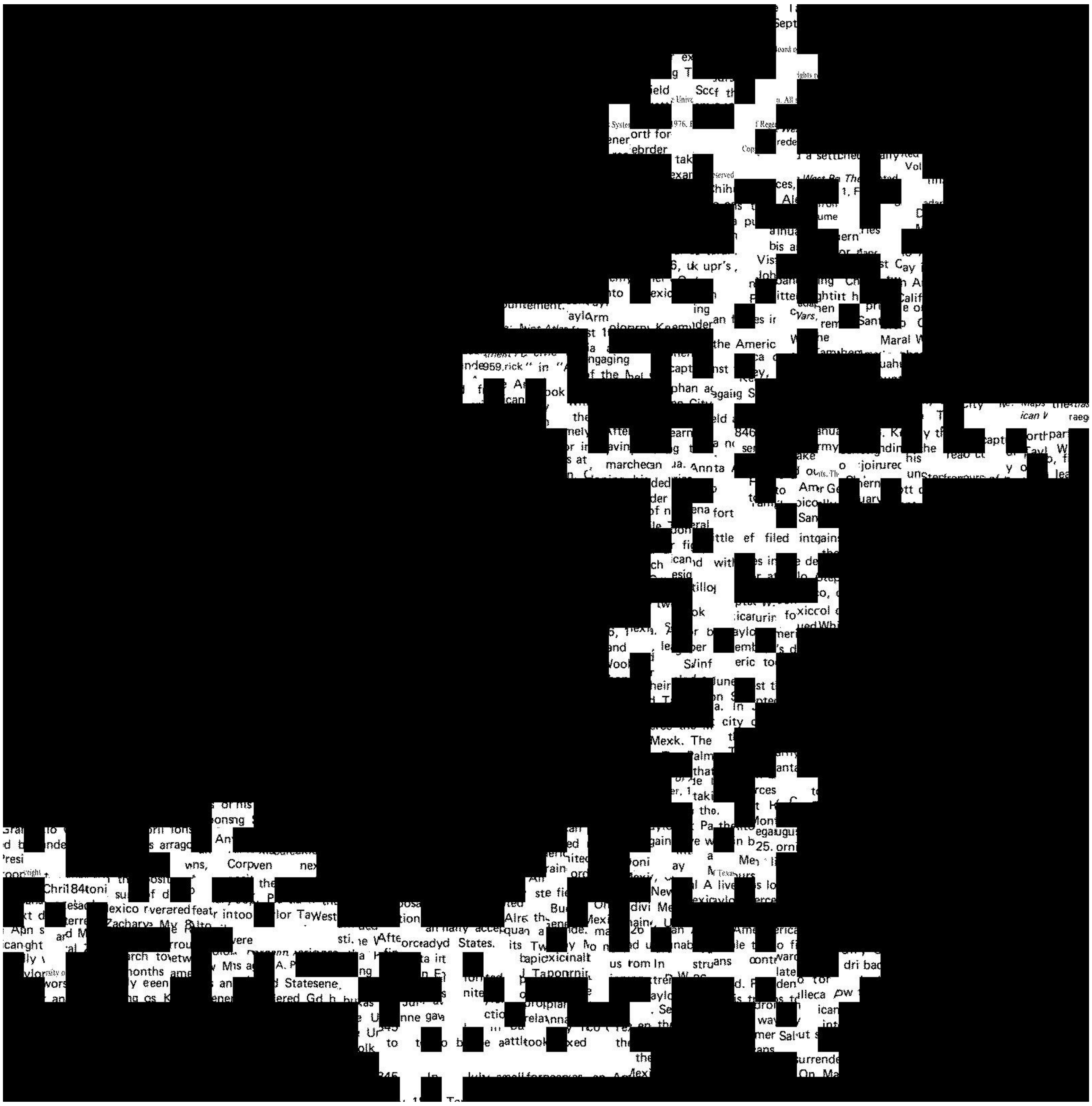
An analysis of various search heuristics also reveals some interesting behaviour . Specifically, when using simple cost functions, there is no cost encountered for the shape of the final document, only for the internal edges. Therefore, a smart search function will manage to “cheat” by minimizing the number of internal edges(picture below). In this situation improving the search function further is pointless until the cost function is fixed to remove this problem. A Bayesian approach to the cost function should again help to ameliorate this problem without introducing any additional overhead of heuristics.

## Conclusion:

I have made a start on the systematic analysis of the nature of unshredding and the costs and benefits of currently proposed solutions. I have also developed a new heuristic which performs on par with all but the most recent solutions in literature.

I plan to continue looking at the application of general machine learning methods such as neural networks and support vector machines and of statistical approaches to the definition of cost and search functions in the hope of formulating more resilient and noise-resistant solutions.

I also plan to test these solutions on real world situations, removing some or all of the convenient but unrealistic assumptions that have been made so far in literature.


**Reconstruction by a good search function using a poor cost function - the search function finds a way to cheat**

## References:

[1]  Matthias Prandtstetter and Gunther R. Raidl - Combining Forces to Reconstruct Strip Shredded Text Documents

[2] Matthias Prandtstetter - Two Approaches for Computing Lower Bounds on the Reconstruction of Strip Shredded Text Documents

[3] Matthias Prandtstetter and Gunther R. Raidl - Meta-Heuristics for Reconstructing Cross Cut Shredded Text Documents

[4]  Christian Schauer, Matthias Prandtstetter, Gunther R. Raidl - A Memetic Algorithm for Reconstructing Cross-Cut Shredded Text Documents

[5]  Azzam Sleit,  Yacoub Massad, Mohammed Musaddaq - An alternative clustering approach for reconstructing cross cut shredded text documents