# Construct Forensic Evidence Networks from Information Fragments

Zhouzhou Li(✉), Xiaoming Liu, Mario Alberto Garcia, and Charles D. McAllister

Southeast Missouri State University, Cape Girardeau, MO 63701, USA
{zli2,xliu,mgarcia,cdmcallister}@semo.edu

**Abstract.** The advancement of modern technologies led to many challenges for digital forensic investigators, who now need to know every fragmentary, trivial, and isolated digital information to stitch an overall, solid evidence picture for each case. While the reality is they lack a mature model to help them accomplish their high-demanded works (such as data collection, protection, retrieval, recovery, analytic, archive). In this paper, we introduce a new concept, "Evidence Networks", and propose to build solid evidence networks from information pieces by reusing the existing successful network models. Additionally, the jigsaw puzzle model is used to analyze the practical problems that an evidence network may encounter.

**Keywords:** Digital forensics · Network model · Path cost · Correlation coefficient

## 1 Introduction

The word forensics means "to bring to court" [1]. It is the application of scientific principles to provide evidence in criminal cases. Its major tasks include how to collect crime scene evidence, prove the causes of accidents, test crime scene evidence in labs, etc. It is a cross-disciplinary science with close connections to crime scene investigation, drug analysis, genetics, physics, organic chemistry, criminal procedures, and the criminal justice system [2].

Digital Forensics (DF) is the modern day version of forensic science and deals with the retrieval, recovery, investigation, and use of material found in digital devices. Using the data collected from electronic devices, digital forensic investigators can assist in recovering lost or stolen data, trace it back to the source, and help create a detailed investigative report that can remedy any crime [3].

Supported by the Institute of Cybersecurity and the Department of Computer Science, Southeast Missouri State University.


J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 702–714, 2021.
https://doi.org/10.1007/978-3-030-89814-4_51

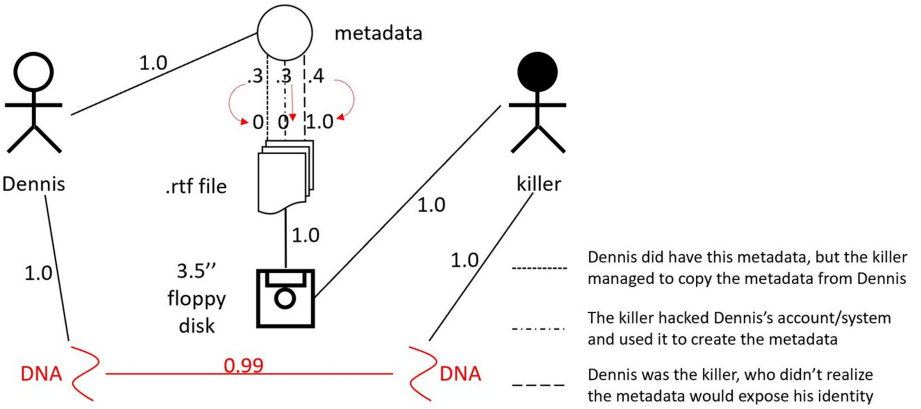


**Fig. 1.** Evidence network for the BTK case

In the famous BTK Killer case, the killer was responsible for the murder of ten people between 1974 and 1991, who was not identified for many years, until he sent a 1.44 MB floppy disk to the police (to mock the police's tracing capability). The digital forensic investigator for this case found a .rtf file on the disk and studied its metadata, which usually is not observable to non-technical users. The author's name and the associated organization left in the metadata provided solid evidence about the identity of the killer. With the guidance of this clue, the killer was eventually arrested and pleaded guilty. This is an good example of constructing solid evidence or clues from the fragmentary, trivial, and isolated digital data, where traditional forensics does not work.

Digital forensics has expanded from computer forensics to cover investigation of all devices capable of storing digital data [4]. With roots in the personal computing revolution of the late 1970s and early 1980s, the discipline evolved in a haphazard manner during the 1990s, and it was not until the early 21st century that national policies emerged.

With more and more technologies (including hardware and software) being introduced to the world, the forms of data storage and transmission changed substantially. The source data for investigation could be found in a PC, in a cell phone, cached in a small cell (base station or access point), or even distributed in multiple physical storage devices (servers) with each piece has no apparent relation to other pieces. Also, plenty data formats (software) for different types of information were invented. The emergence of Internet, Cloud Computing, and IoT, as well as their prevalence, intensifies the situation. For example, a $1.5 IoT device could be equipped with a multiple-core CPU, multiple wireless transceivers, a multiple-MB flash storage and a proprietary file system, which make it a valuable data source for forensic investigation. However, constructing a whole picture from all evidence pieces becomes more difficult.

Not only to deal with different hardware and software for data storage and transmission, a digital forensic investigator also needs to handle incomplete data.

A data user may not be an expert in techniques, who operates the data at higher layer, leaving the physical layer out of his/her attention. For example, a user keeps adding and deleting small files on his/her PC, which leaves many fragments on the hard disk drive. Some fragments may not be cleaned timely, so that they may still contain important and useful information for digital forensics, while without the attentions from the user(s). There is also a possibility that the criminals try to destroy, cover, or tamper the real data for a malicious purpose. A digital forensic investigator needs to grasp all kinds of tools/skills for data recovery.

In short, besides getting familiar with the modern hardware for data storage and transmission, the tools for data retrieval and analysis, constructing solid evidence from the fragmentary, covered, and tampered material is a big challenge to forensic investigators.

Our insight in the course of a digital forensic investigation is the forensics investigators can make full use of the outcomes of modern Computer Science to improve their work efficiency. If we treat all evidence pieces as network nodes, stitching them together is like finding a path from the source node to the destination node. The path cost will be decided by all the segments (hops). Therefore, the existing routing algorithm can be reused for digital forensics.

The remainder of the paper is organized as the follows. In the 'Review' section, we study the existing work for constructing evidence from isolated information. In the 'Proposal' section, we present the "Evidence Network" concept and re-use the existing network models (especially the router models) to describe, analog, and analyze the evidence networks. Also, in this section, we use the classical puzzle stitching problem to analog the problems we may encounter with the evidence network and provide our solution to those practical problems. To demo the counter-intuitive effects for false negative and positive cases, we put quantitative illustrations in the experiment section. In the 'Conclusion and Future Work' section, we conclude our work and indicate how to extend our work in future research.

## 2 Review

To our best knowledge, this work is the first to apply network models and graphic models for forensic evidence stitching.

In [5], the authors listed issues of DF. In [6], the authors explained the potential crisis behind today's golden age of digital forensic due to lacking a clear strategy for enabling research efforts that build upon one another. Especially, they pointed out that data management will be a big challenge to DF. "Even data that can be analyzed can wait weeks or months before review because of data management issues". They listed a few high watched areas for improvement: "the growing size of storage devices", "the increasing prevalence of embedded flash storage and the proliferation of hardware interfaces", "the proliferation of operating systems and file formats", using Cloud for remote processing and storage, and to split a single data structure into elements. All these can be summarized

as a "isolated huge information islands" problem - each piece of information becomes trivial with comparison to the huge storage size, while it has unknown connections to the external.

To solve this problem, previous literature applied different approaches. However, they all have practical issue. Thus, this problem is still not addressed well.

With the success of Internet and Ciscos, network models become matured. They can connect isolated nodes to form a bigger network. The problem that network models solved has highly-abstract similarity with the DF problem we are concerned - just replace the keyword 'node' with 'evidence' in the previous sentence, we will find both have a similar structure.

The essence of a network model has its root in graph theory. Nodes and edges are the elements. In order to find a similar graphic model for our DF problem, we studied different models with focus on "Mosaic", "Panoramic", and "Jigsaw Puzzle" models.

In [7], the authors pointed out that "the format used to represent image data can be as critical in image processing as the algorithms applied to the data". That means, an algorithm cannot cover different formats. Data transforming is a general activity for image comparison and recognition. Then minor difference between the original and transformed images should be expected and not counted as an important factor.

In [8], the passive authentication techniques, operate in the absence of digital watermarks, signatures, or specialized hardware were discussed. This work revealed even there is no direct, obvious connection between two evidence pieces, there are indirect approaches to build/reconstruct their connections.

[9] and [10] introduced how to construct Mosaic. [11] introduced image alignment and image stitching algorithms. Both inspired our ideas for the DF problem.

[12] proposed to create full view panoramic mosaics from image sequences.[13] proposed to automatically create panoramic images based on local and overlap information. As a summary, they tried to utilize the temporal and spacial overlaps to build adjacent relations between pieces. This idea is hard to apply to evidence networks because evidence pieces may not have partial overlaps, but the "subset comparison" idea led us to the "Jigsaw Puzzle" model. [13] also explained their methods for eliminating visible shifts in brightness and removing moving objects which appeared in several image pieces, which gave some hints for us to solve the 'false positive' problem.

[14] revealed the connection between 'global' and 'local': apply global alignment (block adjustment) to the whole sequence of images while warps each image based on the results of pairwise local image registrations. It reminded us to think about the DF problem not just from the local comparison's perspective.

[15–21] proposed practical ideas and algorithms for file carving, which is a process of reassembling computer files from fragments in the absence of file system metadata. These papers used statistical methods or information theory (entropy) to categorize fragments into bins/bags to achieve high classification accuracy. Their methods can be re-used to permute evidences. However, a file

fragment has a fixed size, which is not analog to a evidence because different evidences most likely are heterogeneous. Also, a file fragment only has one prior and one successor. While an evidence may have multiple connections to external. File carving theories and models cannot sufficiently describe high dimension evidence network.

## 3 Proposal

The forensic investigators can collect as many as possible evidence fragments from a criminal scene, the suspect's home, office, and other related places. However, they need a tool or model to piece the fragments into a whole picture of the crime that occurred.

Our proposal is to reuse the network models (especially, the router models) with necessary modifications to describe fragmentary evidences and the correlations between them. We aim to construct a network of evidences and figure out the high correlation path from one person/evidence to another person/evidence, which can guide further investigation.

As shown in Fig. 1, the BTK killer with unknown identity (black head) gave the police a floppy disk. The connection between the floppy disk and the killer has a 1.0 metric, which means a high correlation. The floppy disk contains a .rtf file. The connection between the .rtf file ad the floppy disk also has a 1.0 metric, which means another high correlation. It is worthy to note that (the content of) the .rtf file must be highly correlated to the case, otherwise we cannot give a 1.0 weight to this connection. To clarify this point, we can think about a Microsoft hidden file, which can be seen in almost every floppy disk. But, of course, Microsoft should not be put in the suspect list. The investigator found the name 'Dennis' (as well as his organization) in the metadata of the .rtf file, which causes the high correlation (metric = 1.0) between the metadata and 'Dennis'. However, there is an uncertainty about how the .rtf file can get such a metadata. We consider 3 possibilities here:

1. This 'Dennis' did have this metadata, but the killer managed to copy the metadata from this 'Dennis'
2. The killer hacked 'Dennis' account/system and used it to create the metadata.
3. 'Dennis' was the killer, who didn't realize the metadata would expose his identity.

To represent the uncertainty, we use 3 lines to connect the .rtf file and its metadata. Each line has an initial correlation/weight (0.3, 0.3, and 0.4), with the last one having somewhat higher weight than the others. These are rough estimates. Because at this moment, the investigators do not have detailed information about them, rough figures are fine. We can adjust them later to reflect any new update, just like when updating costs for network links.

Then from the killer to 'Dennis', we can find 3 (non-loop) paths with the following metrics:

- Path 1's metric = 1.0 * 1.0 * 0.3 * 1.0 = 0.3
- Path 2's metric = 1.0 * 1.0 * 0.3 * 1.0 = 0.3
- Path 3's metric = 1.0 * 1.0 * 0.4 * 1.0 = 0.4

The difference between traditional network models and this evidence network model is the way calculate the path 'cost'. Here we use the product of all connections' metrics, while traditional network models use the sum.

We can calculate a sum for all paths. In this case, it is 0.3 + 0.3 + 0.4 = 1, which means a strong correlation between the killer and this 'Dennis', no matter if they are the same person or not. The uncertainty is how they are correlated. Among all the paths, the most 'correlated' path has the metric '0.4', which is the best path so far. Therefore, the strategy would lead to investigate this path first.

In the BTK killer case, later, the investigator determined another path from the killer, who once left his DNA to a victim's nail, to this 'Dennis' - the investigator managed to get 'Dennis' DNA indirectly and found it matched the killer's.

Through this example, we can see how the constructed "Evidence Network" can provide guidance to the digital forensic investigations:

1. Even though there is uncertainty, a strong connection between the killer and 'Dennis' is identified (calculated). The metric for this purpose is the sum of all paths.
2. A best path (like a 'shortest' path in Internet) will be investigated first. To calculate the metric of a path, we use the product of every segment's correlation/weight. The idea is similar to a routing algorithm. The difference is the way we calculate the 'cost' of a path.
3. The correlation between two evidences/persons could be dynamic or initially rough. With some possibilities are proved true or false later, we can update the correlation metric. Just like the path cost fluctuating in Internet. In Fig. 1, (later) after we see the DNAs matching, we can update the correlations between the metadata and the .rtf file from [0.3, 0.3, 0.4] to [0, 0, 1.0].
4. The sum of all paths could be greater than 1.0. This implies that more than one solid evidence has been identified.

Therefore, a simple network topology can be used to describe the BTK killer scenario. We have the killer, the floppy disk, the .rtf file, the metadata, 'Dennis' in the topology. They are the nodes in the evidence network. Like in a real network, there are connections between the nodes, which represent the "correlations" between two nodes. By using the network model, we construct a path from the killer to 'Dennis'. And we use correlations to quantify the path and an overall correlation can be calculated by multiply all segments' correlations (we will discuss why the product rather than the sum is used to describe the overall path).

The BTK killer case is a simple example of 'Evidence Networks'. We need to consider more complicated examples to see if the idea still works.

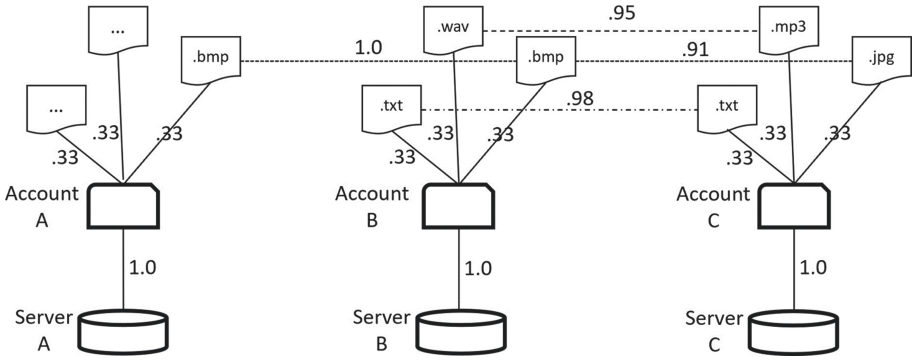


**Fig. 2.** Evidence network for associated accounts

Figure 2 shows an evidence network constructed to help decide whether two accounts are associated. Account B on Server B owns 3 files: a .bmp, a .txt, and a .wav. Account C on Server C also owns 3 files: a .jpg, a .txt, and a .mp3. The C's .mp3 file looks like a compressed version of B's .wav file because by applying a comparison algorithm we find they are similar. Also, the C's .jpg file looks like a compressed version of B's .bmp file because they are similar. The C's .txt file looks like a minor modified version of B's .txt file because they only have a few words different. Without considering Account A, we are pretty sure Account B and C are associated accounts. However, Account A has a .bmp file, which is the same as the .bmp file of B. Because it is a copy of the original .bmp file (not a compressed version), the two .bmp files have a 1.0 correlation, which is the highest correlation in the evidence network. Shall we grant A an associated account? What if we can only put two accounts in the associated account list? A+B or B+C, which is better? This is actually a global-local issue, or multi-dimensional issue.

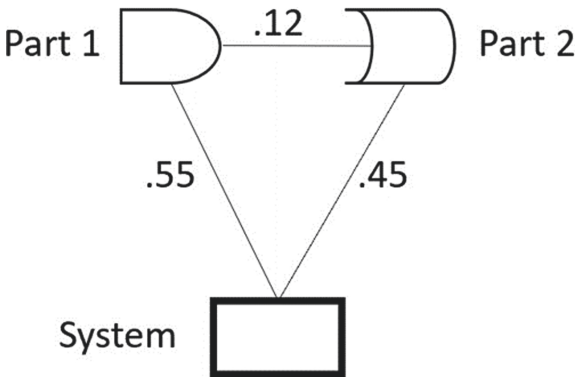


**Fig. 3.** A evidence network with complemented parts

Figure 3 gives another example of 'Evidence Networks'. Part 1 and Part 2 have a very weak correlation (metric = 0.12 if we apply an algorithm to calculate their correlation coefficients) because they are two different parts of a system. However, they should be combined together to form the system. Part 1 has a (metric = 0.55) connection to the system. Part 2 has (metric = 0.45) connection to the system. Using the system as the bridge, Part 1 can reach Part 2 with a 'cost' = 0.55 * 0.45 = 0.2475, which is higher than 0.12. The correlation between the two parts is higher than what we see directly. This is a kind of 'false negative' issue.

There may be other potential issues. We cannot enumerate all cases to derive the potential issues when we apply the 'Evidence Network' model for digital forensic investigation. Therefore, we need an intuitive and powerful tool to identify and analyze all the issues.

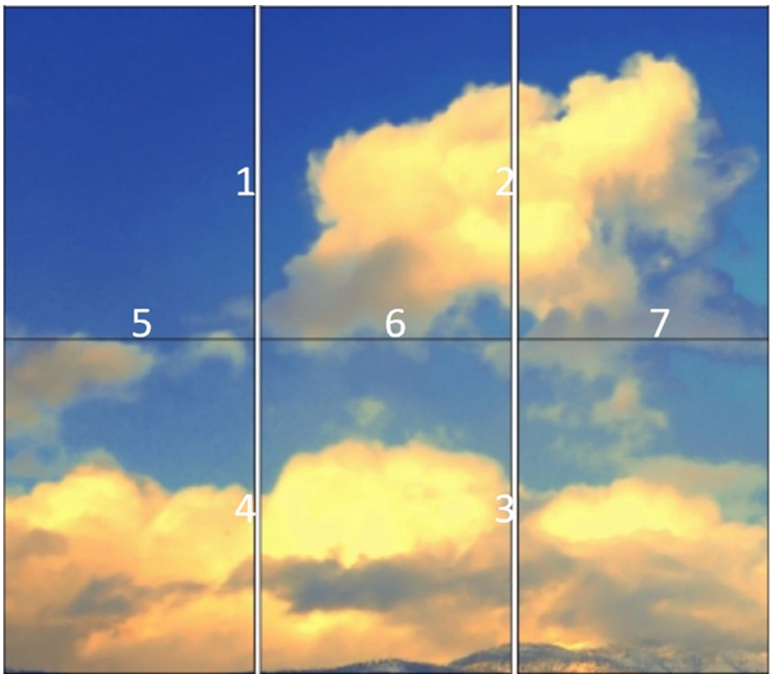


**Fig. 4.** A equal-size-piece puzzle made from [22]

Forming an Evidence Network is analog to stitching all puzzle pieces to from a complete picture. Both rely on local adjacent relations to construct a global graph. Both are expected to have the same issues. Therefore, we use a simple puzzle (Fig. 4) to intuitively explain the possible issues an Evidence Network may encounter.

The basic method to rebuild the original image is to find adjacent pieces for each piece. Comparing the edges of every pair of pieces can help identify the neighbors of a piece.

In Fig. 4, because each piece in the puzzle has the same shape, we can only exploit their color features to conclude their neighbor pieces. Only the four edges of a piece will be compared to the four edges of other pieces to find similar edges, which are candidate adjacent edges. The central area of each piece is ignored for this purpose. In computer graphics, edges are represented by tensors. We can calculate their Pearson Correlation Coefficients to evaluate their similarity. Pearson Correlation Coefficients are values between $[-1.0, 1.0]$. The bigger the absolute value is, the stronger is their similarity. In Fig. 4, two edges indicated by cutting line 2, 3, 4, or 7 have high correlation coefficients (more than 0.89), thus they are similar. Even we shuffle all the pieces, it is easy to determine their adjacent relations in a row through calculation.

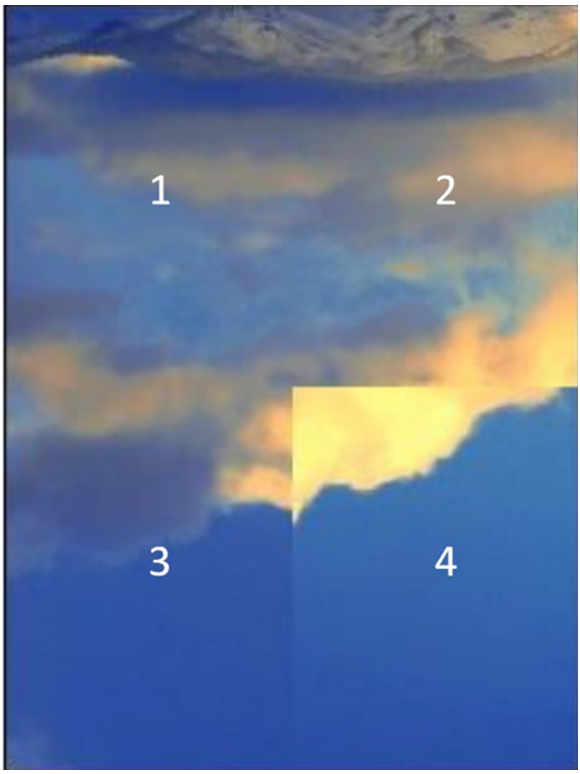


**Fig. 5.** False positive case

However, it is more difficult to determine their adjacent relations in a column. For example, the two edges indicated by cutting line 5 or 6 do not have high

Pearson correlation coefficients. This is because the cutting lines happen to be located around the natural color separation curves. These will be the false negative cases if we only rely on the Pearson correlation. Fortunately, each piece has four edges, i.e., four dimensions. Even when we cannot find a qualified adjacent piece in one dimension, we can try other dimensions. This 'false negative case' issue is equivalent to the 'Complemented Part' issue of the Evidence Networks - the cutting line is located at a natural separation, which makes two pieces/parts not similar at all. The solution is also given by the puzzle example: try other dimensions/paths to globally evaluate the adjacent relation of two pieces/parts.

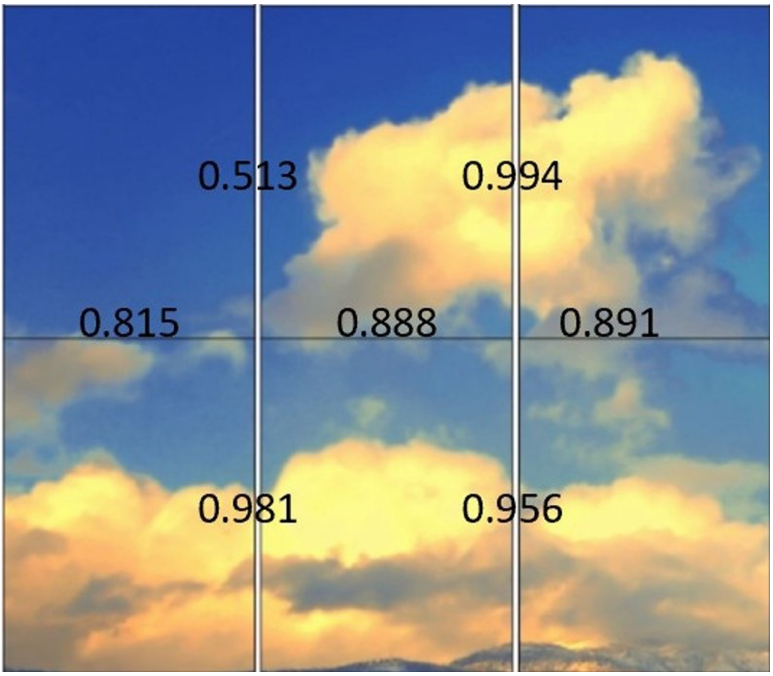


**Fig. 6.** False negative case with counterintuitive quantitative information

Figure 5 gives a 'False Positive' example of the puzzle stitching algorithm. It is obvious that piece 4 should not be there. But, if we compare the edges between piece 3 and 4, we will find they have a high Pearson correlation coefficient. That's the reason why the algorithm will put piece 4 there. Again, this reflects the calculation result from only one dimension. If we consider other edges and use a vote mechanism, piece 4 should be excluded from the candidates. This issue is equivalent to the 'Associated Accounts' issue in the Evidence Networks. Both have one and only one perfectly matching dimension. The solution is: if more than one candidates are found, global and multi-dimensional evaluation should be conducted.

## 4 Experiments

We calculated the correlation coefficients for each adjacent line in Fig. 4 and Fig. 5. Both Fig. 6 and Fig. 7 provide quantitative illustration of the false negative and positive cases, which are counter-intuitive though.

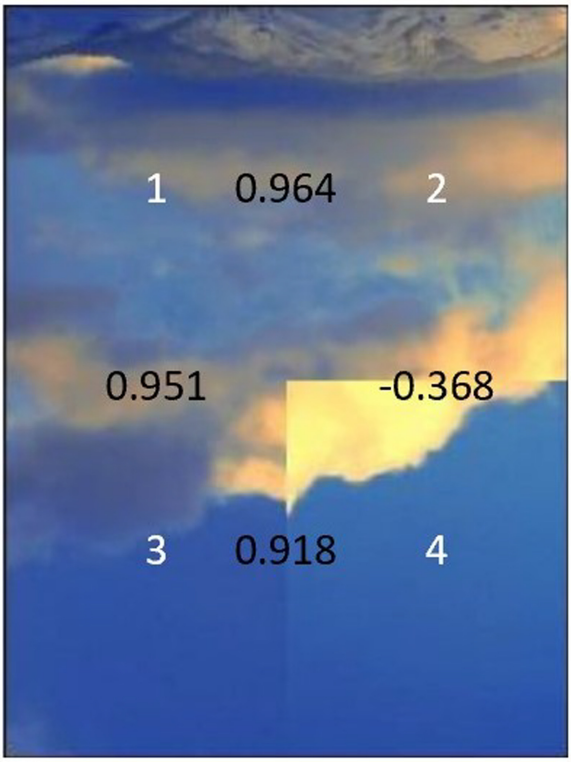


**Fig. 7.** False positive case with counterintuitive quantitative information

## 5 Conclusion and Future Work

In this paper, we reuse the well-developed network models and graphic models for constructing evidence networks for digital forensic investigation. Our evidence network model can reveal the hidden connection between two evidence fragment, the uncertainty of the connection, as well as the guidance to the promising investigation direction. Like a router model, the connection could dynamically vary. Our work also solved the 'false positive' and 'false negative' problems by giving an intuitive explanation from the similar 'Jigsaw Puzzle' model.

In the future, we plan to introduce noise to the models and apply filters to augment the models.

## References


1. Hayes, D.R.: A Practical Guide to Computer Forensics Investigations. Pearson Education, London (2015)
2. https://study.com/articles/Criminology_vs_Criminalistics_Whats_the_Difference.html. Criminology Vs. Criminalistics: What's the Difference? Accessed 8 Jan 2021
3. https://cytelligence.com/resource/what-is-digital-forensics. What Is Digital Forensics? Accessed 8 Jan 2021
4. Reith, M., Carr, C., Gunsch, G.: An examination of digital forensic models. Int. J. Digit. Evid. **1**(3), 1–12 (2002)
5. Delp, E., Memon, N., Min, W.: Digital forensics. IEEE Signal Process. Mag. **26**(2), 14–15 (2009)
6. Garfinkel, S.L.: Digital forensics research: the next 10 years. Digit. Investig. **7**, S64–S73 (2010)
7. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Eng. **29**(6), 33–41 (1984)
8. Farid, H.: Image forensics. Annu. Rev. Vis. Sci. (2019)
9. Hansen, M., Anandan, P., Dana, K., Van der Wal, G., Burt, P.: Real-time scene stabilization and mosaic construction. In: Proceedings of 1994 IEEE Workshop on Applications of Computer Vision, pp. 54–62. IEEE (1994)
10. Burt, P.J., Adelson, E.H.: A multiresolution spline with application to image mosaics. ACM Trans. Graph. (TOG) **2**(4), 217–236 (1983)
11. Szeliski, R.: Image alignment and stitching: a tutorial. Found. Trends® Comput. Graph. Vis. **2**(1), 1–104 (2006)
12. Szeliski, R., Shum, H.-Y.: Creating full view panoramic image mosaics and environment maps. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 251–258 (1997)
13. Uyttendaele, M., Eden, A., Skeliski, R.: Eliminating ghosting and exposure artifacts in image mosaics. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 2, pp. II–II. IEEE (2001)
14. Shum, H.-Y., Szeliski, R.: Systems and experiment paper: construction of panoramic image mosaics with global and local alignment. Int. J. Comput. Vis. **36**(2), 101–130 (2000)
15. Richard III, G.G., Roussev, V.: Scalpel: a frugal, high performance file carver. In: DFRWS (2005)
16. Memon, N., Pal, A.: Automated reassembly of file fragmented images using greedy algorithms. IEEE Trans. Image Process. **15**(2), 385–393 (2006)
17. Garfinkel, S.L.: Carving contiguous and fragmented files with fast object validation. Digit. Investig. **4**, 2–12 (2007)
18. Hand, S., Lin, Z., Guofei, G., Thuraisingham, B.: Bin-carver: automatic recovery of binary executable files. Digit. Investig. **9**, S108–S117 (2012)
19. Casey, E., Zoun, R.: Design tradeoffs for developing fragmented video carving tools. Digit. Investig. **11**, S30–S39 (2014)
20. Qiu, W., Zhu, R., Guo, J., Tang, X., Liu, B., Huang, Z.: A new approach to multimedia files carving. In: 2014 IEEE International Conference on Bioinformatics and Bioengineering, pp. 105–110. IEEE (2014)

21. van der Meer, V., et al.: File Fragmentation in the wild: a privacy-friendly approach. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE (2019)
22. https://www.tmonews.com/2014/11/t-mobile-wideband-lte-touches-down-in-boise/