

METHOD, SYSTEM, AND COMPUTER PROGRAM PRODUCT FOR PROVIDING SYNTHETIC TRANSACTION DATA USING GENERATIVE ARTIFICIAL INTELLIGENCE

BACKGROUND

1. Technical Field

[0001] This disclosure relates generally to detecting anomalies in computing systems and, in some non-limiting embodiments or aspects, to methods, systems, and computer program products for providing synthetic interaction data, such as transaction data, using generative artificial intelligence.

2. Technical Considerations

[0002] Computer simulation may refer to a process of mathematical modeling, performed on a computer, which is designed to predict the behavior of, or the outcome of, a real-world or physical system. A computer simulation may be a useful tool for mathematical modeling of many natural systems in physics (e.g., computational physics), astrophysics, climatology, chemistry, biology, and manufacturing, as well as human systems in economics, psychology, social science, healthcare, and engineering. Computer simulation of a system may be represented as the running of a model of the system. In this way, computer simulation can be used to explore and gain new insights into technology and to estimate the performance of systems that are too complex for analytical solutions.

[0003] In some instances, a computer simulation may be used to determine vulnerabilities and/or recurring patterns with regard to a behavior of a system. For example, a cyber-attack simulation may be used to proactively identify and/or address security gaps within the cybersecurity infrastructure of an organization. Such a simulation can provide a realistic approach to understanding how resilient the cybersecurity infrastructure is to different types of cyber security threats.

[0004] However, modeling the behavior of a system based on real data, such as real transaction data, may lead to the release of sensitive information. Furthermore, simple simulations, which may include the use of simple models and/or tools, may lead to inaccurate data for a computer simulation.

SUMMARY

[0005] Accordingly, provided are improved methods, systems, and computer program products for providing synthetic transaction data using generative artificial intelligence.

[0006] According to non-limiting embodiments or aspects, provided is a computer-implemented method for providing synthetic interaction data, such as transaction data, using generative artificial intelligence. In some non-limiting embodiments or aspects, the computer-implemented method may include receiving, with at least one processor, a dataset including a plurality of data instances. The plurality of data instances may be associated with a plurality of payment transactions involving a plurality of real entities. In some non-limiting embodiments or aspects, the computer-implemented method may further include generating a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances. In some non-limiting embodiments or aspects, generating the first simulation graph may include: generating, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determining an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities; generating a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generating second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances. In some non-limiting embodiments or aspects, the computer-implemented method may further include adding abnormal transaction patterns to the first simulation graph to provide a second simulation graph. In some non-limiting embodiments or aspects, the computer-implemented method may further include training a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0007] In non-limiting embodiments or aspects, the computer-implemented method may further include generating a graph model based on the plurality of payment transactions involving the plurality of real entities. In some non-limiting embodiments or aspects, the computer-implemented method may further include determining a plurality of real subgraphs of the graph model based on transaction behavior of each

real entity of the plurality of real entities. In some non-limiting embodiments or aspects, generating the plurality of simulation subgraphs of the first simulation graph may include generating the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0008] In non-limiting embodiments or aspects, determining the plurality of real subgraphs of the graph model may include: selecting one or more nodes to be included in each real subgraph of the plurality of real subgraphs; clustering the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determining a ratio of each type of real subgraph clusters to all types of real subgraph clusters. In some non-limiting embodiments or aspects, generating the first simulation graph may include generating the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters may be the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0009] In non-limiting embodiments or aspects, adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph may include adding fraudulent transaction patterns to the first simulation graph to provide the second simulation graph. In some non-limiting embodiments or aspects, training the machine learning model may include training a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0010] In non-limiting embodiments or aspects, generating the first simulated transaction data for each simulated entity of the plurality of simulated entities may include providing an input to the one or more first transformer machine learning models to generate an output based on the input. In some non-limiting embodiments or aspects, the input may include a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity. In some non-limiting embodiments or aspects, the output may include a prediction of at least one of: whether a transaction may be sent by the simulated entity within a time interval, whether a transaction may be received by the simulated entity within a time

interval, or whether the simulated entity may be involved in a transaction within a time interval.

[0011] In non-limiting embodiments or aspects, determining the entity type of each simulated entity of the plurality of simulated entities may include: generating an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and clustering each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters may be associated with an entity type of a plurality of entity types.

[0012] In non-limiting embodiments or aspects, generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models may include: generating in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and generating out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0013] According to non-limiting embodiments or aspects, provided is a system for providing synthetic interaction data, such as transaction data, using generative artificial intelligence. In some non-limiting embodiments or aspects, the system may include at least one processor configured to receive a dataset including a plurality of data instances. In some non-limiting embodiments or aspects, the plurality of data instances may be associated with a plurality of payment transactions involving a plurality of real entities. In some non-limiting embodiments or aspects, the at least one processor may be further configured to generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances. In some non-limiting embodiments or aspects, when generating the first simulation graph, the at least one processor may be configured to: generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determine an entity type of each simulated entity of the plurality of simulated entities based on the

first simulated transaction data for each simulated entity of the plurality of simulated entities; generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances. In some non-limiting embodiments or aspects, the at least one processor may be further configured to add abnormal transaction patterns to the first simulation graph to provide a second simulation graph. In some non-limiting embodiments or aspects, the at least one processor may be further configured to train a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0014] In non-limiting embodiments or aspects, the at least one processor may be further configured to generate a graph model based on the plurality of payment transactions involving the plurality of real entities. In some non-limiting embodiments or aspects, the at least one processor may be further configured to determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities. In some non-limiting embodiments or aspects, when generating the plurality of simulation subgraphs of the first simulation graph, the at least one processor may be configured to generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0015] In non-limiting embodiments or aspects, when determining the plurality of real subgraphs of the graph model, the at least one processor may be configured to: select one or more nodes to be included in each real subgraph of the plurality of real subgraphs; cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters. In some non-limiting embodiments or aspects, when generating the first simulation graph, the at least one processor may be configured to generate the first simulation graph to include a plurality of types of simulation subgraphs. In some non-limiting embodiments or aspects, a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters may be the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0016] In non-limiting embodiments or aspects, when adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, the at least one processor may be configured to add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph. In some non-limiting embodiments or aspects, when training the machine learning model, the at least one processor may be configured to train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0017] In non-limiting embodiments or aspects, when generating the first simulated transaction data for each simulated entity of the plurality of simulated entities, the at least one processor may be configured to provide an input to the one or more first transformer machine learning models to generate an output based on the input. In some non-limiting embodiments or aspects, the input may include a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of a real entity. In some non-limiting embodiments or aspects, the output may include a prediction of at least one of: whether a transaction may be sent by the simulated entity within a time interval, whether a transaction may be received by the simulated entity within a time interval, or whether the simulated entity may be involved in a transaction within a time interval.

[0018] In non-limiting embodiments or aspects, when determining the entity type of each simulated entity of the plurality of simulated entities, the at least one processor may be configured to: generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters. In some non-limiting embodiments or aspects, each cluster of the plurality of clusters may be associated with an entity type of a plurality of entity types.

[0019] In non-limiting embodiments or aspects, when generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, the at least one processor may be configured to: generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of

simulation subgraphs using an in-subgraph transformer machine learning model; and generate out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0020] According to non-limiting embodiments or aspects, provided is a computer program product for providing synthetic interaction data, such as transaction data, using generative artificial intelligence. In some non-limiting embodiments or aspects, the computer program product may include at least one non-transitory computer-readable medium including one or more instructions that, when executed by at least one processor, cause the at least one processor to receive a dataset including a plurality of data instances. In some non-limiting embodiments or aspects, the plurality of data instances may be associated with a plurality of payment transactions involving a plurality of real entities. In some non-limiting embodiments or aspects, the one or more instructions may further cause the at least one processor to generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances. In some non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to generate the first simulation graph, may cause the at least one processor to: generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities; generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances. In some non-limiting embodiments or aspects, the one or more instructions may further cause the at least one processor to add abnormal transaction patterns to the first simulation graph to provide a second simulation graph. In some non-limiting embodiments or aspects, the one or more instructions may further cause the at least one processor to train a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0021] In non-limiting embodiments or aspects, the one or more instructions may further cause the at least one processor to generate a graph model based on the plurality of payment transactions involving the plurality of real entities. In some non-limiting embodiments or aspects, the one or more instructions may further cause the at least one processor to determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities. In some non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to generate the plurality of simulation subgraphs of the first simulation graph, may cause the at least one processor to generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0022] In non-limiting embodiments or aspects, when determining the plurality of real subgraphs of the graph model, the one or more instructions may further cause the at least one processor to: select one or more nodes to be included in each real subgraph of the plurality of real subgraphs; cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters. In some non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to generate the first simulation graph, may further cause the at least one processor to generate the first simulation graph to include a plurality of types of simulation subgraphs. In some non-limiting embodiments or aspects, a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters may be the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0023] In non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to add the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, may cause the at least one processor to add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph. In some non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to train the machine learning model, may cause the at least one processor to train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0024] In non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to generate the first simulated transaction data for each simulated entity of the plurality of simulated entities, may cause the at least one processor to provide an input to the one or more first transformer machine learning models to generate an output based on the input. In some non-limiting embodiments or aspects the input may include a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity. In some non-limiting embodiments or aspects, the output may include a prediction of at least one of: whether a transaction may be sent by the simulated entity within a time interval, whether a transaction may be received by the simulated entity within a time interval, or whether the simulated entity may be involved in a transaction within a time interval.

[0025] In non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to determine the entity type of each simulated entity of the plurality of simulated entities, may cause the at least one processor to: generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters. In some non-limiting embodiments or aspects each cluster of the plurality of clusters may be associated with an entity type of a plurality of entity types.

[0026] In non-limiting embodiments or aspects, the one or more instructions that cause the at least one processor to generate the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, may cause the at least one processor to generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and generate out-subgraph simulated transaction data for transactions involving entities in two different subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0027] Other non-limiting embodiments or aspects will be set forth in the following numbered clauses:

[0028] Clause 1: A computer-implemented method, comprising: receiving, with at least one processor, a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities; generating, with at least one processor, a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein generating the first simulation graph comprises: generating, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determining an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities; generating a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generating second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances; adding, with at least one processor, abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and training, with at least one processor, a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0029] Clause 2: The computer-implemented method of clause 1, further comprising: generating a graph model based on the plurality of payment transactions involving the plurality of real entities; determining a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities; wherein generating the plurality of simulation subgraphs of the first simulation graph comprises: generating the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0030] Clause 3: The computer-implemented method of clause 1 or 2, wherein determining the plurality of real subgraphs of the graph model comprises: selecting one or more nodes to be included in each real subgraph of the plurality of real subgraphs; clustering the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determining a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and

wherein generating the first simulation graph comprises: generating the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0031] Clause 4: The computer-implemented method of any of clauses 1-3, wherein adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph comprises: adding fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and wherein training the machine learning model comprises: training a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0032] Clause 5: The computer-implemented method of any of clauses 1-4, wherein generating the first simulated transaction data for each simulated entity of the plurality of simulated entities comprises: providing an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity, and wherein the output comprises a prediction of at least one of: whether a transaction is to be sent by the simulated entity within a time interval, whether a transaction is to be received by the simulated entity within a time interval, or whether the simulated entity will be involved in a transaction within a time interval.

[0033] Clause 6: The computer-implemented method of any of clauses 1-5, wherein determining the entity type of each simulated entity of the plurality of simulated entities comprises: generating an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and clustering each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

[0034] Clause 7: The computer-implemented method of any of clauses 1-6, wherein generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine

learning models comprises: generating in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and generating out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0035] Clause 8: A system, comprising: at least one processor configured to: receive a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities; generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein, when generating the first simulation graph, the at least one processor is configured to: generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities; generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances; add abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and train a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0036] Clause 9: The system of clause 8, wherein the at least one processor is further configured to: generate a graph model based on the plurality of payment transactions involving the plurality of real entities; determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities; and wherein, when generating the plurality of simulation subgraphs of the first simulation graph, the at least one processor is configured to: generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0037] Clause 10: The system of clause 8 or 9, wherein, when determining the plurality of real subgraphs of the graph model, the at least one processor is configured to: select one or more nodes to be included in each real subgraph of the plurality of real subgraphs; cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and wherein, when generating the first simulation graph, the at least one processor is configured to: generate the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0038] Clause 11: The system of any of clauses 8-10, wherein, when adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, the at least one processor is configured to: add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and wherein, when training the machine learning model, the at least one processor is configured to: train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0039] Clause 12: The system of any of clauses 8-11, wherein, when generating the first simulated transaction data for each simulated entity of the plurality of simulated entities, the at least one processor is configured to: provide an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of a real entity, and wherein the output comprises a prediction of at least one of: whether a transaction is to be sent by the simulated entity within a time interval, whether a transaction is to be received by the simulated entity within a time interval, or whether the simulated entity will be involved in a transaction within a time interval.

[0040] Clause 13: The system of any of clauses 8-12, wherein, when determining the entity type of each simulated entity of the plurality of simulated entities, the at least one processor is configured to: generate an embedding for each simulated entity of

the plurality of simulated entities using the one or more first transformer machine learning models; and cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

[0041] Clause 14: The system of any of clauses 8-13, wherein, when generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, the at least one processor is configured to: generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and generate out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0042] Clause 15: A computer program product comprising at least one non-transitory computer-readable medium including one or more instructions that, when executed by at least one processor, cause the at least one processor to: receive a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities; generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein, the one or more instructions that cause the at least one processor to generate the first simulation graph, cause the at least one processor to: generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models; determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities; generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances; add abnormal transaction patterns to the first simulation graph to provide a second

simulation graph; and train a machine learning model based on the second simulation graph to provide a trained machine learning model.

[0043] Clause 16: The computer program product of clause 15, wherein the one or more instructions further cause the at least one processor to: generate a graph model based on the plurality of payment transactions involving the plurality of real entities; determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities; and wherein, the one or more instructions that cause the at least one processor to generate the plurality of simulation subgraphs of the first simulation graph, cause the at least one processor to: generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

[0044] Clause 17: The computer program product of clause 15 or 16, wherein, the one or more instructions that cause the at least one processor to determine the plurality of real subgraphs of the graph model, cause the at least one processor to: select one or more nodes to be included in each real subgraph of the plurality of real subgraphs; cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and wherein, the one or more instructions that cause the at least one processor to generate the first simulation graph, cause the at least one processor to: generate the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0045] Clause 18: The computer program product of any of clauses 15-17, wherein, the one or more instructions that cause the at least one processor to add the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, cause the at least one processor to: add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and wherein, the one or more instructions that cause the at least one processor to train the machine learning model, cause the at least one processor to: train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0046] Clause 19: The computer program product of any of clauses 15-18, wherein, the one or more instructions that cause the at least one processor to generate the first simulated transaction data for each simulated entity of the plurality of simulated entities, cause the at least one processor to: provide an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity, and wherein the output comprises a prediction of at least one of: whether a transaction is to be sent by the simulated entity within a time interval, whether a transaction is to be received by the simulated entity within a time interval, or whether the simulated entity will be involved in a transaction within a time interval.

[0047] Clause 20: The computer program product of any of clauses 15-19, wherein, the one or more instructions that cause the at least one processor to determine the entity type of each simulated entity of the plurality of simulated entities, cause the at least one processor to: generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

[0048] Clause 21: The computer program product of any of clauses 15-20, wherein, the one or more instructions that cause the at least one processor to generate the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, cause the at least one processor to: generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and generate out-subgraph simulated transaction data for transactions involving entities in two different subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0049] These and other features and characteristics of the present disclosure, as well as the methods of operation and functions of the related elements of structures and the combination of parts and economies of manufacture, will become more

apparent upon consideration of the following description and the appended claims with reference to the accompanying drawings, all of which form a part of this specification, wherein like reference numerals designate corresponding parts in the various figures. It is to be expressly understood, however, that the drawings are for the purpose of illustration and description only and are not intended as a definition of the limits of the disclosed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0050] Additional advantages and details are explained in greater detail below with reference to the non-limiting, exemplary embodiments that are illustrated in the accompanying schematic figures, in which:

[0051] FIG. 1 is a schematic diagram of a system for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects;

[0052] FIG. 2 is a flow diagram of a method for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects;

[0053] FIG. 3 is a flow diagram of a method for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects;

[0054] FIGS. 4A-4L are schematic diagrams of an exemplary implementation of a system and/or method for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects;

[0055] FIG. 5 is a diagram of an exemplary environment in which methods, systems, and/or computer program products, described herein, may be implemented, according to some non-limiting embodiments or aspects; and

[0056] FIG. 6 is a schematic diagram of example components of one or more devices of FIG. 1 and/or FIG. 5, according to some non-limiting embodiments or aspects.

DETAILED DESCRIPTION

[0057] For purposes of the description hereinafter, the terms “end,” “upper,” “lower,” “right,” “left,” “vertical,” “horizontal,” “top,” “bottom,” “lateral,” “longitudinal,” and derivatives thereof shall relate to the embodiments as they are oriented in the drawing

figures. However, it is to be understood that the embodiments may assume various alternative variations and step sequences, except where expressly specified to the contrary. It is also to be understood that the specific devices and processes illustrated in the attached drawings, and described in the following specification, are simply exemplary embodiments or aspects of the disclosed subject matter. Hence, specific dimensions and other physical characteristics related to the embodiments or aspects disclosed herein are not to be considered as limiting.

[0058] It is to be understood that the present disclosure may assume various alternative variations and step sequences, except where expressly specified to the contrary. It is also to be understood that the specific devices and processes illustrated in the attached drawings, and described in the following specification, are simply exemplary and non-limiting embodiments or aspects. Hence, specific dimensions and other physical characteristics related to the embodiments or aspects disclosed herein are not to be considered as limiting.

[0059] Some non-limiting embodiments or aspects are described herein in connection with thresholds. As used herein, satisfying a threshold may refer to a value being greater than the threshold, more than the threshold, higher than the threshold, greater than or equal to the threshold, less than the threshold, fewer than the threshold, lower than the threshold, less than or equal to the threshold, equal to the threshold, etc.

[0060] No aspect, component, element, structure, act, step, function, instruction, and/or the like used herein should be construed as critical or essential unless explicitly described as such. Also, as used herein, the articles “a” and “an” are intended to include one or more items and may be used interchangeably with “one or more” and “at least one.” Furthermore, as used herein, the term “set” is intended to include one or more items (e.g., related items, unrelated items, a combination of related and unrelated items, and/or the like) and may be used interchangeably with “one or more” or “at least one.” Where only one item is intended, the term “one” or similar language is used. Also, as used herein, the terms “has,” “have,” “having,” or the like are intended to be open-ended terms. Further, the phrase “based on” is intended to mean “based at least partially on” unless explicitly stated otherwise. In addition, reference to an action being “based on” a condition may refer to the action being “in response to” the condition. For example, the phrases “based on” and “in response to” may, in some non-limiting embodiments or aspects, refer to a condition for automatically triggering

an action (e.g., a specific operation of an electronic device, such as a computing device, a processor, and/or the like).

[0061] As used herein, the term “communication” may refer to the reception, receipt, transmission, transfer, provision, and/or the like of data (e.g., information, signals, messages, instructions, commands, and/or the like). For one unit (e.g., a device, a system, a component of a device or system, combinations thereof, and/or the like) to be in communication with another unit means that the one unit is able to directly or indirectly receive information from and/or transmit information to the other unit. This may refer to a direct or indirect connection (e.g., a direct communication connection, an indirect communication connection, and/or the like) that is wired and/or wireless in nature. Additionally, two units may be in communication with each other even though the information transmitted may be modified, processed, relayed, and/or routed between the first and second unit. For example, a first unit may be in communication with a second unit even though the first unit passively receives information and does not actively transmit information to the second unit. As another example, a first unit may be in communication with a second unit if at least one intermediary unit processes information received from the first unit and communicates the processed information to the second unit. In some non-limiting embodiments or aspects, a message may refer to a network packet (e.g., a data packet and/or the like) that includes data. It will be appreciated that numerous other arrangements are possible.

[0062] As used herein, the term “computing device” may refer to one or more electronic devices configured to process data. A computing device may, in some examples, include the necessary components to receive, process, and output data, such as a processor, a display, a memory, an input device, a network interface, and/or the like. A computing device may be a mobile device. As an example, a mobile device may include a cellular phone (e.g., a smartphone or standard cellular phone), a portable computer, a wearable device (e.g., watches, glasses, lenses, clothing, and/or the like), a personal digital assistant (PDA), and/or other like devices. A computing device may also be a desktop computer or other form of non-mobile computer.

[0063] As used herein, the term “server” may refer to or include one or more computing devices that are operated by or facilitate communication and processing for multiple parties in a network environment, such as the Internet, although it will be appreciated that communication may be facilitated over one or more public or private

network environments and that various other arrangements are possible. Further, multiple computing devices (e.g., servers, point-of-sale (POS) devices, mobile devices, etc.) directly or indirectly communicating in the network environment may constitute a “system.”

[0064] As used herein, the term “system” may refer to one or more computing devices or combinations of computing devices (e.g., processors, servers, client devices, software applications, components of such, and/or the like). Reference to “a device,” “a server,” “a processor,” and/or the like, as used herein, may refer to a previously-recited device, server, or processor that is recited as performing a previous step or function, a different device, server, or processor, and/or a combination of devices, servers, and/or processors. For example, as used in the specification and the claims, a first device, a first server, or a first processor that is recited as performing a first step or a first function may refer to the same or different device, server, or processor recited as performing a second step or a second function.

[0065] Non-limiting embodiments or aspects of the disclosed subject matter are directed to systems, methods, and computer program products for providing synthetic data using generative artificial intelligence. Non-limiting embodiments or aspects of the disclosed subject matter provide for an AI simulation system that is configured to receive a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities, generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, add abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and train a machine learning model, such as a fraud detection machine learning model based on the second simulation graph to provide a trained machine learning model.

[0066] In some non-limiting embodiments, when generating the first simulation graph the AI simulation system may generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models, determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities, generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities, and

generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances.

[0067] In some non-limiting embodiments, the AI simulation system may be configured to generate a graph model based on the plurality of payment transactions involving the plurality of real entities and determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities. In some non-limiting embodiments, when generating the plurality of simulation subgraphs of the first simulation graph, the AI simulation system may be configured to generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model. In some non-limiting embodiments, when generating the plurality of real subgraphs of the graph model, the AI simulation system may be configured to select one or more nodes to be included in each real subgraph of the plurality of real subgraphs, cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters, and determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters. In some non-limiting embodiments, when generating the first simulation graph, the AI simulation system may be configured to generate the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

[0068] In some non-limiting embodiments, when adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, the AI simulation system may be configured to add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph. In some non-limiting embodiments, when training the machine learning model, the AI simulation system may be configured to train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

[0069] In some non-limiting embodiments, when generating the first simulated transaction data for each simulated entity of the plurality of simulated entities, the AI simulation system may be configured to provide an input to the one or more first

transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity, and wherein the output comprises a prediction of at least one of: whether a transaction is to be sent by the simulated entity within a time interval, whether a transaction is to be received by the simulated entity within a time interval, or whether the simulated entity will be involved in a transaction within a time interval.

[0070] In some non-limiting embodiments, when determining the entity type of each simulated entity of the plurality of simulated entities, the AI simulation system may be configured to generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models and cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

[0071] In some non-limiting embodiments, when generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, the AI simulation system may be configured to generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model and generate out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

[0072] In this way, non-limiting embodiments or aspects of the disclosed subject matter provide for modeling the behavior of a system (e.g., a payment processing system, such as a transaction service provider system) based on real data, such as real transaction data, that prevents the release of sensitive information. Furthermore, by using complex machine learning models, such as transformer machine learning models, the disclosed subject matter provides for more accurate results of a computer simulation as compared to simpler modeling techniques.

[0073] For the purpose of illustration, in the following description, while the presently disclosed subject matter is described with respect to methods, systems, and

computer program products for providing synthetic data using generative artificial intelligence, such as for use during processing of data associated with interactions (e.g., such as transactions), one skilled in the art will recognize that the disclosed subject matter is not limited to the non-limiting embodiments or aspects disclosed herein. For example, the methods, systems, and computer program products, described herein, may be used with a wide variety of settings, such as providing synthetic data using generative artificial intelligence in any suitable setting, e.g., predictions, regressions, classifications, fraud prevention, authorization, authentication, identification, feature selection, and/or the like.

[0074] Referring now to FIG. 1, shown is a diagram of system 100 for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects. System 100 may include artificial intelligence (AI) simulation system 102, machine learning (ML) model management repository 104, user device 106, and/or communication network 108. In some non-limiting embodiments or aspects, AI simulation system 102, ML model management repository 104, and/or user device 106 may interconnect (e.g., establish a connection to communicate) via wired connections, wireless connections, or a combination of wired and wireless connections.

[0075] AI simulation system 102 may include one or more devices that are capable of being in communication with ML model management repository 104 and/or user device 106 via communication network 108. For example, AI simulation system 102 may include a computing device, such as a server (e.g., a single server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, AI simulation system 102 may include a processor and/or memory, as described herein. In some non-limiting embodiments or aspects, AI simulation system 102 may include one or more software instructions (e.g., one or more software applications) executing on a server (e.g., a single server), a group of servers, a computing device (e.g., a single computing device), a group of computing devices, and/or other like devices. In some non-limiting embodiments or aspects, AI simulation system 102 may be configured to perform one or more steps of methods described herein. In some non-limiting embodiments or aspects, AI simulation system 102 may be configured to communicate with a data storage device (e.g., ML model management repository 104). In some non-limiting embodiments or aspects, AI simulation system 102 may be in communication with ML model management repository 104 and/or user device

106, such that AI simulation system 102 is separate from ML model management repository 104 and/or user device 106. In some non-limiting embodiments or aspects, ML model management repository 104 and/or user device 106 may be implemented by (e.g., may be part of) AI simulation system 102.

[0076] Additionally or alternatively, AI simulation system 102 may generate (e.g., train, validate, re-train, and/or the like), store, and/or implement (e.g., operate, provide inputs to and/or outputs from, and/or the like) one or more machine learning models. For example, AI simulation system 102 may generate one or more machine learning models by fitting (e.g., validating) one or more machine learning models against data used for training (e.g., training data). In some non-limiting embodiments or aspects, AI simulation system 102 may generate, store, and/or implement one or more machine learning models that are provided for a real-time environment (e.g., a runtime environment) used for providing inferences based on data in a live situation. In some non-limiting embodiments or aspects, AI simulation system 102 may be in communication with a data storage device (ML model management repository 104), which may be local or remote to AI simulation system 102.

[0077] ML model management repository 104 may include one or more devices that are capable of being in communication with AI simulation system 102 and/or user device 106 via communication network 108. For example, ML model management repository 104 may include a computing device, such as a server (e.g., a single server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, ML model management repository 104 may be associated with one or more computing devices providing interfaces, such that a user (e.g., an administrative user, a user using a service account, and/or the like) may interact with ML model management repository 104 via the one or more computing devices. ML model management repository 104 may be in communication with AI simulation system 102 and/or user device 106, such that ML model management repository 104 is separate from AI simulation system 102 and/or user device 106. Alternatively, in some non-limiting embodiments or aspects, ML model management repository 104 may be implemented by (e.g., may be part of) AI simulation system 102 and/or user device 106.

[0078] User device 106 may include one or more devices that are capable of being in communication with AI simulation system 102 and/or ML model management repository 104 via communication network 108. For example, user device 106 may

include a computing device, such as a desktop computer, a portable computer (e.g., tablet computer, a laptop computer, and/or the like), a mobile device (e.g., a cellular phone, a smartphone, a personal digital assistant, a wearable device, and/or the like), and/or other like devices. In some non-limiting embodiments or aspects, user device 106 may be associated with a user (e.g., an individual operating user device 106).

[0079] Communication network 108 may include one or more wired and/or wireless networks. For example, communication network 108 may include a cellular network (e.g., a long-term evolution (LTE) network, a third-generation (3G) network, a fourth-generation (4G) network, a fifth-generation (5G) network, a code division multiple access (CDMA) network, etc.), a public land mobile network (PLMN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), a telephone network (e.g., the public switched telephone network (PSTN) and/or the like), a private network, an ad hoc network, an intranet, the Internet, a fiber optic-based network, a cloud computing network, and/or the like, and/or a combination of some or all of these or other types of networks.

[0080] The number and arrangement of systems and devices shown in FIG. 1 are provided as an example. There may be additional systems and/or devices, fewer systems and/or devices, different systems and/or devices, and/or differently arranged systems and/or devices than those shown in FIG. 1. Furthermore, two or more systems or devices shown in FIG. 1 may be implemented within a single system or device, or a single system or device shown in FIG. 1 may be implemented as multiple distributed systems or devices. Additionally or alternatively, a set of systems (e.g., one or more systems) or a set of devices (e.g., one or more devices) of system 100 may perform one or more functions described as being performed by another set of systems or another set of devices of system 100.

[0081] Referring now to FIG. 2, shown is a flow diagram of process 200 for providing synthetic transaction data using generative artificial intelligence, according to some non-limiting embodiments or aspects. The steps shown in FIG. 2 are for example purposes only. It will be appreciated that additional, fewer, different, and/or a different order of steps may be used in non-limiting embodiments or aspects. In some non-limiting embodiments or aspects, a step may be automatically performed in response to performance and/or completion of a prior step.

[0082] In some non-limiting embodiments or aspects, one or more of the steps of process 200 may be performed (e.g., completely, partially, etc.) by AI simulation

system 102 (e.g., one or more devices of AI simulation system 102). In some non-limiting embodiments or aspects, one or more of the steps of process 200 may be performed (e.g., completely, partially, etc.) by another device or a group of devices separate from or including AI simulation system 102, such as ML model management repository 104 (e.g., one or more devices of ML model management repository 104), and/or user device 106.

[0083] As shown in FIG. 2, at step 202, process 200 may include receiving a dataset that includes a plurality of data instances associated with a plurality of transactions. For example, AI simulation system 102 may receive the dataset that includes the plurality of data instances from ML model management repository 104, user device 106, and/or another system or device. In some non-limiting embodiments, the plurality of data instances of the dataset may be associated with a plurality of real transactions (e.g., non-simulated transactions, such as payment transactions that have been processed by a transaction service provider) involving a plurality of real entities (e.g., non-simulated entities, such as consumers, account holders, etc., that conduct payment transactions).

[0084] In some non-limiting embodiments or aspects, the dataset may include a plurality of data instances (e.g., data records), where each data instance has a label. For example, the dataset may include a plurality of data instances, where each data instance has a label indicating whether the data instance is associated with a non-fraudulent data instance (e.g., a data instance associated with a non-fraudulent payment transaction) or a fraudulent data instance (e.g., a data instance associated with fraudulent payment transaction). In some non-limiting embodiments or aspects, the dataset includes a plurality of data records associated with a plurality of payment transactions, wherein each data record has a label indicating whether the data record is associated with a payment transaction that is fraudulent or associated with a payment transaction that is non-fraudulent.

[0085] In some non-limiting embodiments or aspects, a data instance (e.g., a data instance of a dataset, such as a training dataset) may include transaction data associated with a payment transaction involving a real entity. In some non-limiting embodiments or aspects, the transaction data may include a plurality of transaction parameters associated with an electronic payment transaction. In some non-limiting embodiments or aspects, the plurality of features may represent the plurality of transaction parameters. In some non-limiting embodiments or aspects, the plurality of

transaction parameters may include electronic wallet card data associated with an electronic card (e.g., an electronic credit card, an electronic debit card, an electronic loyalty card, and/or the like), decision data associated with a decision (e.g., a decision to approve or deny a transaction authorization request), authorization data associated with an authorization response (e.g., an approved spending limit, an approved transaction value, and/or the like), a primary account number (PAN), an authorization code (e.g., a personal identification number (PIN), etc.), data associated with a transaction amount (e.g., an approved limit, a transaction value, etc.), data associated with a transaction date and time, data associated with a conversion rate of a currency, data associated with a merchant type (e.g., a merchant category code that indicates a type of goods, such as grocery, fuel, and/or the like), data associated with an acquiring institution country, data associated with an identifier of a country associated with the PAN, data associated with a response code, data associated with a merchant identifier (e.g., a merchant name, a merchant location, and/or the like), data associated with a type of currency corresponding to funds stored in association with the PAN, and/or the like.

[0086] In some examples, the dataset may include a large amount of data instances, such as 100 data instances, 500 data instances, 1,000 data instances, 5,000 data instances, 10,000 data instances, 25,000 data instances, 50,000 data instances, 100,000 data instances, 1,000,000 data instances, and/or the like. In some non-limiting embodiments or aspects, a percentage (e.g., a first percentage) of the plurality of data instances are labeled with a first label (e.g., labeled with a positive label of a binary classification, labeled with a negative label of a binary classification, etc.). In some non-limiting embodiments or aspects, the plurality of data instances are labeled based on labels provided as an output of a deep learning model (e.g., a deep learning fraud detection model).

[0087] As shown in FIG. 2, at step 204, process 200 may include generating a first simulation graph representing a plurality of simulated transactions among a plurality of simulated entities. For example, AI simulation system 102 may generate the first simulation graph (e.g., an initial or basic simulation graph) based on the dataset (e.g., the dataset that includes a plurality of data instances associated with a plurality of transactions involving a plurality of real entities).

[0088] In some non-limiting embodiments or aspects, AI simulation system 102 may generate a graph model (e.g., a mathematical structure used to model pairwise

relations between objects) based on the plurality of transactions involving the plurality of real entities. In some non-limiting embodiments or aspects, the graph model may include a plurality of nodes (e.g., vertices, points, etc.), which are connected by a plurality of edges (e.g., links, lines, etc.), where the nodes represent each real entity of a plurality of real entities and the edges represent transactions (e.g., simulated real-time payment transactions), between the plurality of real entities. The plurality of real entities may include entities involved in a transaction. For example, the plurality of entities may include an entity that uses an account involved in a transaction, such as an account holder (e.g., a card holder, such as a credit card holder, a debit card holder, etc.), an entity that provides goods and/or services for a transaction, such as a merchant, an entity that provides the account (e.g., that issues the account) involved in a transaction, an entity that processes the transaction, such as a transaction service provider, any combination thereof, and/or the like. In some non-limiting embodiments or aspects, each edge of the plurality of edges may represent a data instance of the plurality of data instances. For example, the edges may represent a transaction and may include transaction data associated with the transaction.

[0089] In some non-limiting embodiments or aspects, AI simulation system 102 may generate a graph model based on the plurality of data instances of the dataset. For example, AI simulation system 102 may generate a graph model based on the plurality of payment transactions involving the plurality of real entities. In some non-limiting embodiments or aspects, AI simulation system 102 may determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities. For example, AI simulation system 102 may select one or more nodes (e.g., one or more nodes having a particular transaction behavior) to be included in each (e.g., each type of) real subgraph of a plurality of (e.g., a plurality of types of) real subgraphs. In some non-limiting embodiments or aspects, a real subgraph may have the same or a different number of nodes as another real subgraph. In the example above, AI simulation system 102 may calculate a value of similarity (e.g., cosine similarity) between each real subgraph and all other real subgraphs of the plurality of real subgraphs and may cluster each real subgraph into a real type of subgraph cluster of a plurality of real types of subgraph clusters based on a threshold value of similarity (e.g., a threshold value of 0.8) between each real subgraph cluster. In some non-limiting embodiments or aspects, AI simulation system 102 may cluster all real subgraphs to a real type of subgraph cluster of a plurality of real types of

subgraph clusters. In the example above, AI simulation system 102 may determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters.

[0090] In some non-limiting embodiments or aspects, the first simulation graph may be based on (e.g., may be generated based on) the graph model. For example, AI simulation system 102 may generate the first simulation graph to include a plurality of types of simulation subgraphs, where a ratio of each type of simulation subgraph (e.g., type of simulation subgraph cluster) to all types of simulation subgraphs is the same as a ratio of each type of real subgraph (e.g., type of real subgraph cluster) to all types of real subgraphs. In some non-limiting embodiments or aspects, the first simulation graph may be based on (e.g., consist of) a plurality of simulation subgraphs. Each simulation subgraph may be defined based on transaction behavior of simulated entities included in the simulation subgraph (e.g., transactions between simulated entities included in the simulation subgraph). In some non-limiting embodiments or aspects, each subgraph may be defined based on a number of transactions between simulated entities included in the simulation subgraph, defined based on a type of transaction between simulated entities included in the simulation subgraph, defined based on a number of simulated entities included in the simulation subgraph, an entity type (e.g., an entity type as defined by account behavior of the simulated entity) of each simulated entity included in the simulation subgraph, a number (e.g., an allotted number) of entity types included in the simulation subgraph, and/or any combination thereof.

[0091] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the first simulation graph using one or more transformer machine learning models (e.g., a plurality of transformer machine learning models, such as a first transformer machine learning model, a second transformer machine learning model, and/or a third transformer machine learning model, etc.). For example, AI simulation system 102 may generate the first simulation graph by providing data associated with the plurality of data instances of the dataset as inputs to the one or more transformer machine learning models (e.g., one or more generative artificial intelligence models), and the outputs of the one or more transformer machine learning models may be combined to provide the first simulation graph.

[0092] In some non-limiting embodiments or aspects, the one or more transformer machine learning models may include a neural network machine learning model. For example, the one or more transformer machine learning models may include a neural

network machine learning model having a large number of nodes. In such an example, the one or more transformer machine learning models may include a neural network machine learning model having 10 nodes, 20 nodes, 50 nodes, 100 nodes, 1,000 nodes, and/or the like. Further details regarding generation of the first simulation graph are provided herein (e.g., with regard to FIG. 3).

[0093] As shown in FIG. 2, at step 206, process 200 may include generating a second simulation graph that includes at least one abnormal transaction pattern. For example, AI simulation system 102 may generate the second simulation graph (e.g., a completed or complex simulation graph) that includes at least one abnormal transaction pattern, based on a first simulation graph. In some non-limiting embodiments or aspects, AI simulation system 102 may add transaction patterns (e.g., abnormal transaction patterns, such as fraudulent transaction patterns) to the first simulation graph to provide the second simulation graph. According to some examples, the transaction patterns may include abnormal transaction patterns, which may include a pattern associated with an account take over (ATO), a pattern associated with an authorized push payment (APP) or an unauthorized push payment (UAPP), a pattern associated with a business email compromise (BEC), a pattern associated with a pyramid scheme (e.g., a pyramid scam), and/or any combination thereof.

[0094] As shown in FIG. 2, at step 208, process 200 may include training a machine learning model on the second simulation graph. For example, AI simulation system 102 may train a machine learning model on the second simulation graph to provide a trained machine learning model (e.g., a trained deep learning model). In some non-limiting embodiments or aspects, AI simulation system 102 may train a fraud classification machine learning model (e.g., a fraud prevention model) based on the second simulation graph to provide a trained fraud classification machine learning model.

[0095] In some non-limiting embodiments or aspects, AI simulation system 102 may perform an action in real time using the trained fraud classification machine learning model. For example, AI simulation system 102 may perform an action based on a label (e.g., a classification label) of an input provided by the trained fraud classification machine learning model as an output, wherein the input comprises a data instance associated with a payment transaction conducted in real time. In some non-limiting embodiments or aspects, AI simulation system 102 may perform a

procedure associated with protection of an account of a user (e.g., a user associated with user device 106) based on the label of the input. For example, if the label of the input indicates that the procedure is necessary, AI simulation system 102 may perform the procedure associated with protection of the account of the user. In such an example, if the label of the input indicates that the procedure is not necessary, AI simulation system 102 may forego performing the procedure associated with protection of the account of the user. In some non-limiting embodiments or aspects, AI simulation system 102 may execute a fraud protection procedure based on the label of the input.

[0096] Referring now to FIG. 3, shown is a flow diagram of process 300 for generating a simulation graph (e.g., a first simulation graph, an initial simulation graph, a basic simulation graph, etc.) representing a plurality of simulated transactions among a plurality of simulated entities, according to some non-limiting embodiments or aspects. The steps shown in FIG. 3 are for example purposes only. It will be appreciated that additional, fewer, different, and/or a different order of steps may be used in non-limiting embodiments or aspects. In some non-limiting embodiments or aspects, a step may be automatically performed in response to performance and/or completion of a prior step.

[0097] In some non-limiting embodiments or aspects, one or more of the steps of process 300 may be performed (e.g., completely, partially, etc.) by AI simulation system 102 (e.g., one or more devices of AI simulation system 102). In some non-limiting embodiments or aspects, one or more of the steps of process 300 may be performed (e.g., completely, partially, etc.) by another device or a group of devices separate from or including AI simulation system 102, such as ML model management repository 104 (e.g., one or more devices of ML model management repository 104), and/or user device 106. In some non-limiting embodiments or aspects, one or more of the steps of process 300 may be the same as or similar to step 204 of process 200.

[0098] As shown in FIG. 3, at step 302, process 300 may include generating first simulated transaction data for each simulated entity of a plurality of simulated entities. For example, AI simulation system 102 may generate the first simulated transaction data (e.g., basic simulated transaction data) for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models, such as an entity encoding model (e.g., a many-to-1 seq2seq transformer machine

learning model), an entity behavior generation model (e.g., a many-to-many seq2seq transformer machine learning model), and/or a combination thereof.

[0099] In some non-limiting embodiments or aspects, the entity encoding model may include a trained model designed to provide information regarding transaction behavior of each simulated entity of a plurality of simulated entities in order to determine an entity type of a simulated entity. In some non-limiting embodiments or aspects, the entity encoding model may be configured to receive, as inputs, a plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity may include a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity and/or a plurality of aggregated transaction features associated with a last transaction of the real entity. In some non-limiting embodiments or aspects, the entity encoding model may be configured to provide, as outputs, a prediction of at least one of whether a transaction is to be sent by the simulated entity within a time interval, whether a transaction is to be received by the simulated entity within a time interval, or whether the simulated entity will be involved in a transaction within a time interval.

[0100] In some non-limiting embodiments or aspects, the entity behavior generation model may include a trained model designed to provide information regarding transaction behavior of a simulated entity according to an entity type of the simulated entity. In some non-limiting embodiments or aspects, the entity behavior generation model may be configured to receive, as inputs, features associated with transaction behavior of a real entity corresponding to an entity type of a simulated entity and/or a feature associated with an indication of whether a last transaction of the real entity was sent to or received by the real entity. The account behavior encoding machine learning model may be configured to provide, as outputs, probabilities with regard to whether a transaction is to be sent by the entity type of the simulated entity within a time interval, whether a transaction is to be received by the entity type of the simulated entity within a time interval, and/or whether the entity type of the simulated entity will be involved in a transaction within a time interval.

[0101] In some non-limiting embodiments or aspects, the entity encoding model and/or the entity behavior generation model may provide a state (e.g., a hidden state) that includes an embedding (e.g., a vector that represents an embedding, an

embedding vector, etc.) for each simulated entity (e.g., each simulated entity that corresponds to a real entity). In some non-limiting embodiments or aspects, the first simulated transaction data for each simulated entity may include the embedding for each simulated entity.

[0102] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the first simulated transaction data for each simulated entity based on the dataset (e.g., the plurality of data instances of the dataset). In some non-limiting embodiments or aspects, AI simulation system 102 may generate the first simulated transaction data for each simulated entity by providing an input to the entity encoding model to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving an entity and/or a plurality of aggregated transaction features associated with a last transaction of the entity. In some non-limiting embodiments or aspects, the output comprises a prediction of one of whether a transaction is to be sent by the entity within a time interval (e.g., a time interval of a number of minutes or hours, such as a 10 minute time interval, a 30 minute time interval, a 1 hour time interval, a 5 hour time interval, etc., or a time interval of a number of days, weeks, or months, such as a time interval of 1 day, a time interval of 3 days, a time interval of 1 week, a time interval of 1 month, etc.), whether a transaction is to be received by the entity within a time interval, or (e.g., exclusive or) whether the entity will be involved in a transaction within a time interval. In some non-limiting embodiments or aspects, AI simulation system 102 may generate an embedding for each simulated entity based on a state of the entity encoding model that corresponds to the input provided to the entity encoding model.

[0103] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the plurality of features that are to be provided as inputs to the one or more first transformer machine learning models. For example, AI simulation system 102 may generate the plurality of features based on a plurality of data instances of a dataset. In some non-limiting embodiments or aspects, the plurality of features may include a number of features to increase the accuracy of an output of the one or more first transformer machine learning models. In some examples, the plurality of features that are to be provided as inputs may include 5 features, 10 features, 20 features, 50 features, 100 features, 1,000 features, and/or the like.

[0104] In some non-limiting embodiments or aspects, the plurality of features may include a plurality of time related transaction features of a transaction involving a real

entity, a plurality of aggregated transaction features associated with a last transaction of a real entity, and/or a feature associated with an indication of whether a last transaction of the real entity was sent to or received by the real entity. For example, the plurality of features may include a feature associated with a time of day of a transaction (e.g., an hour of a day of a transaction), a feature associated with a day of a week (e.g., a weekday or a day of a weekend, a specific day of a week, etc.) of a transaction, a feature associated with a week of a year (e.g., an enumerated week out of 52 weeks in a year) of a transaction, a feature associated with an indication of whether a transaction involved a holiday (e.g., a federal holiday), a feature associated with an amount of time between a time of a transaction and a time of a last transaction involving a real entity, a feature associated with an amount of time between a time of a last transaction involving a real entity and a time at which the feature is generated, one or more features associated with a last transaction involving a real entity (e.g., a transaction time, a transaction amount, an incoming transaction, an outgoing transaction, etc.), and/or any combination thereof.

[0105] As shown in FIG. 3, at step 304, process 300 may include determining an entity type of each simulated entity of the plurality of simulated entities. For example, AI simulation system 102 may determine an entity type of each simulated entity of the plurality of simulated entities based on the one or more first transformer machine learning models. In some non-limiting embodiments or aspects, AI simulation system 102 may determine the entity type of each simulated entity of the plurality of simulated entities by generating an embedding for each simulated entity of the plurality of simulated entities (e.g., an embedding for each simulated entity based on a state of the entity encoding model that corresponds to the input provided to the entity encoding model). In some non-limiting embodiments or aspects, AI simulation system 102 may generate the embedding for each simulated entity using the one or more first transformer machine learning models (e.g., by retrieving a state of the one or more first transformer machine learning models).

[0106] In some non-limiting embodiments or aspects, AI simulation system 102 may cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters of simulated entities. Each cluster of the plurality of clusters may be associated with an entity type of a plurality of entity types.

[0107] As shown in FIG. 3, at step 306, process 300 may include generating a plurality of simulation subgraphs of the first simulation graph. For example, AI simulation system 102 may generate the plurality of simulation subgraphs of the first simulation graph based on the dataset of the plurality of data instances. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model. For example, AI simulation system 102 may generate the plurality of simulation subgraphs so that parameters of the plurality of simulation subgraphs are the same as or similar to parameters of the plurality of real subgraphs. In such an example, transaction parameters of transactions involving simulated entities of a simulation subgraph may be the same as or similar to transaction parameters of transactions involving entities of a real subgraph, a number of nodes of a simulation subgraph may be the same as or similar to a number of nodes of a real subgraph, a ratio of entity types of simulated entities of a simulation subgraph may be the same as or similar to a ratio of entity types of entities of a real subgraph, and/or the like.

[0108] In some non-limiting embodiments or aspects, AI simulation system 102 may generate a plurality of simulation subgraphs of the graph model of the first simulation graph based on an entity type of each simulated entity and/or transaction behavior of each simulated entity of the plurality of simulated entities. Additionally or alternatively, AI simulation system 102 may generate the plurality of simulation subgraphs of the first simulation graph based on a plurality of real subgraphs of the graph model. For example, AI simulation system 102 may generate the plurality of simulation subgraphs of the first simulation graph so that a ratio of each entity type of a simulated entity in each simulation subgraph is the same as a ratio of each entity type of a real entity in each real subgraph.

[0109] In some non-limiting embodiments or aspects, AI simulation system 102 may cluster the plurality of simulation subgraphs into a plurality of simulation subgraph clusters (e.g., a plurality of types of subgraph clusters) based on a measure of similarity between simulation subgraphs (e.g., between simulated entities that are included in the simulation subgraphs). For example, AI simulation system 102 may select one or more simulated entities to be included in each simulation subgraph of the plurality of simulation subgraphs and cluster the plurality of simulation subgraphs into a plurality of types of simulation subgraph clusters based on a measure of

similarity between the plurality of types of simulation subgraphs to provide a plurality of simulation subgraph clusters. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the first simulation graph to include a plurality of types of simulation subgraphs, where a ratio of each type of simulation subgraph cluster to all types of simulation subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters in a graph model.

[0110] In some non-limiting embodiments or aspects, AI simulation system 102 may assign a group of simulated entities (e.g., of the plurality of simulated entities) to a simulation subgraph in a simulation subgraph cluster of the plurality of simulation subgraph clusters of the first simulation graph based on an entity type of each simulated entity of the group of simulated entities and an allotted number of entity types for each simulation subgraph in a subgraph cluster. For example, AI simulation system 102 may assign a first group of simulated entities to each simulation subgraph in a first subgraph cluster of the plurality of simulation subgraph clusters based on an entity type of each simulated entity of the first group of simulated entities and/or an allotted number of entity types for each simulation subgraph in a first subgraph cluster, and AI simulation system 102 may assign a second group of simulated entities of the plurality of simulated entities to each simulation subgraph in a second subgraph cluster of the plurality of subgraph clusters based on an entity type of each simulated entity of the second group of simulated entities and/or an allotted number of entity types for each simulation subgraph in a second subgraph cluster.

[0111] As shown in FIG. 3, at step 308, process 300 may include generating second simulated transaction data for each simulated entity of the plurality of simulated entities. For example, AI simulation system 102 may generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, such as an in-subgraph behavior generation model (e.g., an in-subgraph transformer machine learning model) and/or an out-subgraph behavior generation model (e.g., an out-subgraph transformer machine learning model). In some non-limiting embodiments or aspects, AI simulation system 102 may generate second simulated transaction data for each simulated entity of the plurality of simulated entities based on the plurality of data instances of the dataset. In some non-limiting embodiments or aspects, the second simulated transaction data for each simulated entity of the plurality of simulated entities may include synthetic transaction data that is the same as or similar

to transaction data associated with a payment transaction involving a real entity, such that the transaction data may include a plurality of transaction parameters associated with an electronic payment transaction, however, the transaction data will not include sensitive information and will be different from real transaction data (e.g., include values that are different from real transaction data). In some non-limiting embodiments or aspects, the first simulation graph may include the second simulated transaction data for each simulated entity of the plurality of simulated entities.

[0112] In some non-limiting embodiments or aspects, the in-subgraph behavior generation model may include a trained model designed to provide information regarding transaction behavior involving simulated entities of a plurality of simulated entities in a simulation subgraph (e.g., a simulation subgraph of a simulation subgraph cluster) of the first simulation graph. In some non-limiting embodiments or aspects, the in-subgraph behavior generation model may be configured to receive, as inputs, a plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity may include a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity, a feature associated with an amount of a last transaction of the real entity, and/or a state (e.g., a state represented as a vector) of previous transaction activity of the real entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of the real entity may include a plurality of features with regard to transaction sequences according to each real subgraph of the graph model. In some non-limiting embodiments or aspects, the in-subgraph behavior generation model may be configured to provide, as outputs, a prediction of a time at which a transaction will be performed (e.g., an amount of time between a transaction and a last transaction) involving the simulated entity and/or a prediction of an amount of a transaction that will be performed involving the simulated entity. In some non-limiting embodiments or aspects, the in-subgraph behavior generation model may be configured based on a specific real subgraph (e.g., a specific type of real subgraph) of a plurality of real subgraphs or a specific real subgraph cluster of a plurality of real subgraph clusters of the graph model.

[0113] In some non-limiting embodiments or aspects, the out-subgraph behavior generation model may include a trained model designed to provide information

regarding transaction behavior involving simulated entities in different simulation subgraphs of the first simulation graph. In some non-limiting embodiments or aspects, the out-subgraph behavior generation model may be configured to receive, as inputs, a plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity may include a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity, a feature associated with an amount of a last transaction of the real entity, and/or a state (e.g., a state represented as a vector) of previous transaction activity of the real entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of the real entity may include a plurality of features with regard to transaction sequences between real entities of different real subgraphs of the graph model. In some non-limiting embodiments or aspects, the out-subgraph behavior generation model may be configured to provide, as outputs, a prediction of a time at which a transaction will be performed (e.g., an amount of time between a transaction and a last transaction) involving the simulated entity and/or a prediction of an amount of a transaction that will be performed involving the simulated entity. In some non-limiting embodiments or aspects, the out-subgraph behavior generation model may be configured based on a number of different real subgraphs of the graph model.

[0114] In some non-limiting embodiments or aspects, the second simulated transaction data for each simulated entity of the plurality of simulated entities may include in-subgraph simulated transaction data and/or out-subgraph simulated transaction data. In some non-limiting embodiments or aspects, AI simulation system 102 may generate in-subgraph simulated transaction data (e.g., synthetic transaction data that represents transactions involving simulated entities in a same simulation subgraph of the first simulation graph) for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using the in-subgraph behavior generation model and/or generate out-subgraph simulated transaction data (e.g., synthetic transaction data that represents transactions involving simulated entities in different subgraphs of the first simulation graph) for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using the out-subgraph behavior generation model.

[0115] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the second simulated transaction data based on the out-subgraph behavior generation model and the entity behavior generation model. For example, AI simulation system 102 may combine outputs of the out-subgraph behavior generation model and the entity behavior generation model to provide a portion of the second simulated transaction (e.g., out-subgraph simulated transaction data). Additionally or alternatively, AI simulation system 102 may generate the second simulated transaction data based on the in-subgraph behavior generation model and the entity behavior generation model. For example, AI simulation system 102 may combine outputs of the in-subgraph behavior generation model and the entity behavior generation model to provide a portion of the second simulated transaction (e.g., in-subgraph simulated transaction data).

[0116] In some non-limiting embodiments or aspects, to provide the in-subgraph simulated transaction data, AI simulation system 102 may generate a sequence transaction using the in-subgraph behavior generation model for a simulation subgraph, where each transaction has a transaction amount and a timestamp. In some non-limiting embodiments or aspects, AI simulation system 102 may provide the sequence transaction as an input to the entity behavior generation model, and the entity behavior generation model may provide, as an output, probabilities with regard to whether a transaction is to be sent by an entity type of a simulated entity within a time interval, whether a transaction is to be received by an entity type of a simulated entity within a time interval, and/or whether an entity type of a simulated entity will be involved in a transaction within a time interval. For each of the probabilities, AI simulation system 102 may modify the probability by an adjustable parameter for the simulation subgraph to provide final probabilities. In some non-limiting embodiments or aspects, the adjustable parameter may be based on a ratio of entity types of a simulation subgraph in a simulation subgraph cluster, where the adjustable parameter of an entity type can be defined by the following equation:

$$\text{[0117]} \quad -p(\text{entity type}) = -decay^n * \frac{1}{\log(ratio)}$$

[0118] where decay is equal to 0.99, n is a count of an entity type of a simulated entity (e.g., an entity type of a simulated entity that is selected, and the ratio is a percentage of an entity type in a simulation subgraph cluster). In some non-limiting embodiments or aspects, AI simulation system 102 may select an entity type with a

highest probability in sending a transaction and an entity type with a highest probability in receiving a transaction as the pair of entities for a transaction and the transaction may be assigned to the pair of accounts. The steps may be repeated until all transactions are assigned to a pair of entities of the first simulation graph for the in-subgraph simulated transaction data.

[0119] In some non-limiting embodiments or aspects, to provide the out-subgraph simulated transaction data, AI simulation system 102 may randomly select a first simulation subgraph as a sender and a second simulation subgraph as a receiver from the plurality of simulation subgraphs. For each of the first and second simulation subgraphs, and given a timestamp (e.g., as an input), AI simulation system 102 may use the entity behavior generation model to provide, as an output, probabilities with regard to whether a transaction is to be sent by an entity type of a simulated entity within a time interval, whether a transaction is to be received by an entity type of a simulated entity within a time interval, and/or whether an entity type of a simulated entity will be involved in a transaction within a time interval. In some non-limiting embodiments or aspects, AI simulation system 102 may select an entity type with a highest probability in sending a transaction from the first simulation subgraph and an entity type with a highest probability in receiving a transaction from the second simulation subgraph as the pair of entities for a transaction and the transaction may be assigned to the pair of accounts. The steps may be repeated until all transactions are assigned to a pair of entities of the first simulation graph for the out-subgraph simulated transaction data.

[0120] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the second simulated transaction data for each simulated entity based on the dataset (e.g., the plurality of data instances of the dataset). In some non-limiting embodiments or aspects, AI simulation system 102 may generate the second simulated transaction data for each simulated entity by providing an input to the one or more second transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity, a feature associated with an amount of a last transaction of the real entity, and/or a state (e.g., a state represented as a vector) of previous transaction activity of the real entity. In some non-limiting embodiments or aspects, the output comprises a prediction of a time at which a transaction will be performed (e.g., an amount of time between a

transaction and a last transaction, such as an amount of time in seconds) involving the simulated entity and/or a prediction of an amount of a transaction that will be performed involving the simulated entity.

[0121] In some non-limiting embodiments or aspects, AI simulation system 102 may generate the plurality of features that are to be provided as inputs to the one or more second transformer machine learning models. For example, AI simulation system 102 may generate the plurality of features based on a plurality of data instances of a dataset. In some non-limiting embodiments or aspects, the plurality of features may include a number of features to increase the accuracy of an output of the one or more second transformer machine learning models. In some examples, the plurality of features that are to be provided as inputs may include 5 features, 10 features, 20 features, 50 features, 100 features, 1,000 features, and/or the like.

[0122] In some non-limiting embodiments or aspects, the plurality of features may include a plurality of time related transaction features of a transaction involving a real entity, a plurality of aggregated transaction features associated with a last transaction of a real entity, and/or a feature associated with an amount of a last transaction of a real entity. For example, the plurality of features may include a feature associated with a time of day of a transaction (e.g., an hour of a day of a transaction), a feature associated with a day of a week (e.g., a weekday or a day of a weekend, a specific day of a week, etc.) of a transaction, a feature associated with a week of a year (e.g., an enumerated week out of 52 weeks in a year) of a transaction, a feature associated with an indication of whether a transaction involved a holiday (e.g., a federal holiday), a feature associated with an amount of time between a time of a transaction and a time of a last transaction involving a real entity, a feature associated with an amount of time between a time of a last transaction involving a real entity and a time at which the feature is generated, one or more features associated with a last transaction involving a real entity (e.g., a transaction time, a transaction amount, an incoming transaction, an outgoing transaction, etc.), and/or any combination thereof.

[0123] Referring now to FIGS. 4A-4L, FIGS. 4A-4L are diagrams of a non-limiting embodiment or aspect of implementation 400 relating to a process (e.g., process 200 and process 300) for providing synthetic transaction data using generative artificial intelligence. In some non-limiting embodiments or aspects, one or more of the steps of the process may be performed (e.g., completely, partially, etc.) by AI simulation system 102 (e.g., one or more devices of AI simulation system 102). In some non-

limiting embodiments or aspects, one or more of the steps of the process may be performed (e.g., completely, partially, etc.) by another device or a group of devices separate from or including AI simulation system 102 (e.g., one or more devices of AI simulation system 102), ML model management repository 104 (e.g., one or more devices of ML model management repository 104), and/or user device 106.

[0124] As shown by reference number 405 in FIG. 4A, AI simulation system 102 may generate a graph model based on a dataset. In some non-limiting embodiments or aspects, the dataset may include a plurality of data instances that are associated with a plurality of real transactions involving a plurality of real entities. In some non-limiting embodiments or aspects, the graph model may include a plurality of nodes, which are connected by a plurality of edges, where the nodes represent each real entity of a plurality of real entities and the edges represent transactions between real entities of the plurality of real entities.

[0125] As shown by reference number 410 in FIG. 4A, AI simulation system 102 may determine a plurality of (e.g., a plurality of types of) real subgraphs of the graph model. In some non-limiting embodiments or aspects, AI simulation system 102 may determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities. For example, AI simulation system 102 may select one or more nodes (e.g., one or more nodes having a particular transaction behavior) to be included in each (e.g., each type of) real subgraph of a plurality of real subgraphs.

[0126] As shown by reference number 415 in FIG. 4B, AI simulation system 102 may determine a plurality of types of real subgraph clusters of the graph model. In some non-limiting embodiments or aspects, AI simulation system 102 may calculate a value of similarity (e.g., cosine similarity) between each real subgraph and all other real subgraphs of the plurality of real subgraphs and may cluster each real subgraph into a real type of subgraph cluster of a plurality of real types of subgraph clusters based on a threshold value of similarity (e.g., a threshold value of 0.8) between each real subgraph cluster. In some non-limiting embodiments or aspects, AI simulation system 102 may cluster all of the real subgraphs to a plurality of real types of subgraph clusters (e.g., a first real type of subgraph cluster, shown as “sub1”, a second real type of subgraph cluster, shown as “sub2”, a third real type of subgraph cluster, shown as “sub3”, and a fourth real type of subgraph cluster, shown as “sub4”). In some non-

limiting embodiments or aspects, AI simulation system 102 may determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters.

[0127] As shown by reference number 420 in FIG. 4C, AI simulation system 102 may generate a first portion of first simulated transaction data for each simulated entity of the plurality of simulated entities using an entity encoding model. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the first portion of the first simulated transaction data for each simulated entity by providing an input to the entity encoding model to generate an output based on the input, wherein the input may include a plurality of time related transaction features of a transaction involving an entity and a plurality of aggregated transaction features associated with a last transaction of the entity. In some non-limiting embodiments or aspects, the output comprises a prediction of one of whether a transaction is to be sent by the entity within a time interval, whether a transaction is to be received by the entity within a time interval, or (e.g., exclusive or) whether the entity will be involved in a transaction within a time interval. As further shown in FIG. 4C, AI simulation system 102 may generate an embedding for each simulated entity based on a state of the entity encoding model that corresponds to the input provided to the entity encoding model to provide a plurality of embeddings for the plurality of simulated entities.

[0128] As shown by reference number 425 in FIG. 4D, AI simulation system 102 may determine an entity type of each simulated entity of the plurality of simulated entities. In some non-limiting embodiments or aspects, AI simulation system 102 may cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on the plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters of simulated entities. Each cluster of the plurality of clusters may be associated with an entity type of a plurality of entity types, and AI simulation system 102 may assign an entity type to a simulated entity based on the entity type of the cluster of simulated entities to which the simulated entity belongs.

[0129] As shown by reference number 430 in FIG. 4E, AI simulation system 102 may generate a plurality of simulation subgraphs of a first simulation graph. In some non-limiting embodiments or aspects, AI simulation system 102 may generate a plurality of simulation subgraphs to include a plurality of types of simulation subgraphs (e.g., a first type of simulation subgraph, a second type of simulation subgraph, and a third type of simulation subgraph) where a ratio of each type of simulation subgraph

to all types of simulation subgraphs is the same as a ratio of each type of real subgraph to all types of real subgraphs of the graph model. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the plurality of simulation subgraphs, such that a ratio of entity types of simulated entities of each simulation subgraph is the same as a ratio of entity types of real entities of each real subgraph. In some non-limiting embodiments or aspects, the plurality of simulation subgraphs may be clustered into types, such that the plurality of simulation subgraphs include a plurality of types of simulation subgraph clusters.

[0130] As shown by reference number 435 in FIG. 4F, AI simulation system 102 may distribute simulated entities to the plurality of simulation subgraphs based on entity types. In some non-limiting embodiments or aspects, for each type of simulation subgraph cluster, AI simulation system 102 may map simulated entities into simulation subgraphs of a type of simulation subgraph cluster based on entity types.

[0131] As shown by reference number 440 in FIG. 4G, AI simulation system 102 may generate a second portion of first simulated transaction data for each simulated entity of the plurality of simulated entities using an entity behavior generation model. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the second portion of the first simulated transaction data for each simulated entity by providing an input to the entity behavior generation model to generate an output based on the input, wherein the input may include a plurality of time related transaction features associated with transaction behavior of a real entity corresponding to an entity type of a simulated entity and a feature associated with an indication of whether a last transaction of the real entity was sent to or received by the real entity. In some non-limiting embodiments or aspects, the output comprises probabilities with regard to whether a transaction is to be sent by the entity type of the simulated entity within a time interval, whether a transaction is to be received by the entity type of the simulated entity within a time interval, and/or whether the entity type of the simulated entity will be involved in a transaction within a time interval.

[0132] As shown by reference number 445 in FIG. 4H, AI simulation system 102 may generate in-subgraph simulated transaction data using an in-subgraph behavior generation model. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the in-subgraph simulated transaction data for each simulated entity of a simulation subgraph by providing an input to in-subgraph behavior generation model to generate an output based on the input, wherein the input may

include a plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity may include a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity, a feature associated with an amount of a last transaction of the real entity, and/or a state (e.g., a state represented as a vector) of previous transaction activity of the real entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of the real entity may include a plurality of features with regard to transaction sequences according to each real subgraph of the graph model. In some non-limiting embodiments or aspects, the output comprises a prediction of a time at which a transaction will be performed (e.g., an amount of time between a transaction and a last transaction) involving the simulated entity and/or a prediction of an amount of a transaction that will be performed involving the simulated entity.

[0133] As shown by reference number 450 in FIG. 4I, AI simulation system 102 may generate out-subgraph simulated transaction data using an out-subgraph behavior generation model. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the out-subgraph simulated transaction data for a pair of simulated entities of different simulation subgraphs by providing an input to out-subgraph behavior generation model to generate an output based on the input, wherein the input may include a plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of a real entity corresponding to a simulated entity may include a plurality of time related transaction features of a transaction involving the real entity associated with the simulated entity, a feature associated with an amount of a last transaction of the real entity, and/or a state (e.g., a state represented as a vector) of previous transaction activity of the real entity. In some non-limiting embodiments or aspects, the plurality of features associated with transaction behavior of the real entity may include a plurality of features with regard to transaction sequences between real entities of different real subgraphs of the graph model. In some non-limiting embodiments or aspects, the output comprises a prediction of a time at which a transaction will be performed (e.g., an amount of time between a transaction and a last transaction)

involving the simulated entity and/or a prediction of an amount of a transaction that will be performed involving the simulated entity.

[0134] As shown by reference number 455 in FIG. 4J, AI simulation system 102 may generate the first simulation graph. As further shown in FIG. 4J, AI simulation system 102 may generate the first simulation graph based on combining the in-subgraph simulated transaction data, the out-subgraph simulated transaction data, and the first simulated transaction data (e.g., a second portion of the first simulated transaction data).

[0135] As shown by reference number 460 in FIG. 4K, AI simulation system 102 may generate a second simulation graph that includes at least one abnormal transaction pattern. In some non-limiting embodiments or aspects, AI simulation system 102 may generate the second simulation graph by adding a plurality of abnormal transaction patterns to the first simulation graph to provide the second simulation graph. As shown by reference number 465 in FIG. 4L, AI simulation system 102 may train a fraud detection model based on the second simulation graph.

[0136] Referring now to FIG. 5, FIG. 5 is a diagram of a non-limiting embodiment or aspect of an exemplary environment 500 in which systems, products, and/or methods, as described herein, may be implemented. As shown in FIG. 5, environment 500 may include transaction service provider system 502, issuer system 504, customer device 506, merchant system 508, acquirer system 510, and communication network 512. In some non-limiting embodiments or aspects, each of AI simulation system 102, ML model management repository 104, and/or user device 106 may be implemented by (e.g., part of) transaction service provider system 502. In some non-limiting embodiments or aspects, at least one of AI simulation system 102, ML model management repository 104, and/or user device 106 may be implemented by (e.g., part of) another system, another device, another group of systems, or another group of devices, separate from or including transaction service provider system 502, such as issuer system 504, customer device 506, merchant system 508, acquirer system 510, and/or the like.

[0137] Transaction service provider system 502 may include one or more devices capable of receiving information from and/or communicating information to issuer system 504, customer device 506, merchant system 508, and/or acquirer system 510 via communication network 512. For example, transaction service provider system 502 may include a computing device, such as a server (e.g., a transaction processing

server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, transaction service provider system 502 may be associated with a transaction service provider, as described herein. In some non-limiting embodiments or aspects, transaction service provider system 502 may be in communication with a data storage device, which may be local or remote to transaction service provider system 502. In some non-limiting embodiments or aspects, transaction service provider system 502 may be capable of receiving information from, storing information in, communicating information to, or searching information stored in the data storage device.

[0138] Issuer system 504 may include one or more devices capable of receiving information and/or communicating information to transaction service provider system 502, customer device 506, merchant system 508, and/or acquirer system 510 via communication network 512. For example, issuer system 504 may include a computing device, such as a server, a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, issuer system 504 may be associated with an issuer institution, as described herein. For example, issuer system 504 may be associated with an issuer institution that issued a credit account, debit account, credit card, debit card, and/or the like, to a user associated with customer device 506.

[0139] Customer device 506 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 502, issuer system 504, merchant system 508, and/or acquirer system 510 via communication network 512. Additionally or alternatively, each customer device 506 may include a device capable of receiving information from and/or communicating information to other customer devices 506 via communication network 512, another network (e.g., an ad hoc network, a local network, a private network, a virtual private network, and/or the like), and/or any other suitable communication technique. For example, customer device 506 may include a client device and/or the like. In some non-limiting embodiments or aspects, customer device 506 may or may not be capable of receiving information (e.g., from merchant system 508 or from another customer device 506) via a short-range wireless communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like), and/or communicating information (e.g., to merchant system 508) via a short-range wireless communication connection.

[0140] Merchant system 508 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 502, issuer system 504, customer device 506, and/or acquirer system 510 via communication network 512. Merchant system 508 may also include a device capable of receiving information from customer device 506 via communication network 512, a communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like) with customer device 506, and/or the like, and/or communicating information to customer device 506 via communication network 512, the communication connection, and/or the like. In some non-limiting embodiments or aspects, merchant system 508 may include a computing device, such as a server, a group of servers, a client device, a group of client devices, and/or other like devices. In some non-limiting embodiments or aspects, merchant system 508 may be associated with a merchant, as described herein. In some non-limiting embodiments or aspects, merchant system 508 may include one or more client devices. For example, merchant system 508 may include a client device that allows a merchant to communicate information to transaction service provider system 502. In some non-limiting embodiments or aspects, merchant system 508 may include one or more devices, such as computers, computer systems, and/or peripheral devices capable of being used by a merchant to conduct a transaction with a user. For example, merchant system 508 may include a point-of-sale (POS) device and/or a point-of-sale (POS) system.

[0141] Acquirer system 510 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 502, issuer system 504, customer device 506, and/or merchant system 508 via communication network 512. For example, acquirer system 510 may include a computing device, a server, a group of servers, and/or the like. In some non-limiting embodiments or aspects, acquirer system 510 may be associated with an acquirer, as described herein.

[0142] Communication network 512 may include one or more wired and/or wireless networks. For example, communication network 512 may include a cellular network (e.g., a long-term evolution (LTE®) network, a third generation (3G) network, a fourth generation (4G) network, a fifth generation (5G) network, a code division multiple access (CDMA) network, and/or the like), a public land mobile network (PLMN), a local

area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), a telephone network (e.g., the public switched telephone network (PSTN)), a private network (e.g., a private network associated with a transaction service provider), an ad hoc network, an intranet, the Internet, a fiber optic-based network, a cloud computing network, and/or the like, and/or a combination of these or other types of networks.

[0143] The number and arrangement of systems, devices, and/or networks shown in FIG. 5 are provided as an example. There may be additional systems, devices, and/or networks; fewer systems, devices, and/or networks; different systems, devices, and/or networks; and/or differently arranged systems, devices, and/or networks than those shown in FIG. 5. Furthermore, two or more systems or devices shown in FIG. 5 may be implemented within a single system or device, or a single system or device shown in FIG. 5 may be implemented as multiple, distributed systems or devices. Additionally or alternatively, a set of systems (e.g., one or more systems) or a set of devices (e.g., one or more devices) of environment 500 may perform one or more functions described as being performed by another set of systems or another set of devices of environment 500.

[0144] Referring now to FIG. 6, shown is a diagram of example components of a device 600, according to non-limiting embodiments or aspects. Device 600 may correspond to at least one of AI simulation system 102, ML model management repository 104, and/or user device 106 in FIG. 1 and/or at least one of transaction service provider system 502, issuer system 504, customer device 506, merchant system 508, and/or acquirer system 510 in FIG. 5, as an example. In some non-limiting embodiments or aspects, such systems or devices in FIG. 1 or FIG. 5 may include at least one device 600 and/or at least one component of device 600. The number and arrangement of components shown in FIG. 6 are provided as an example. In some non-limiting embodiments or aspects, device 600 may include additional components, fewer components, different components, or differently arranged components than those shown in FIG. 6. Additionally or alternatively, a set of components (e.g., one or more components) of device 600 may perform one or more functions described as being performed by another set of components of device 600.

[0145] As shown in FIG. 6, device 600 may include bus 602, processor 604, memory 606, storage component 608, input component 610, output component 612, and communication interface 614. Bus 602 may include a component that permits communication among the components of device 600. In some non-limiting

embodiments or aspects, processor 604 may be implemented in hardware, firmware, or a combination of hardware and software. For example, processor 604 may include a processor (e.g., a central processing unit (CPU), a graphics processing unit (GPU), an accelerated processing unit (APU), etc.), a microprocessor, a digital signal processor (DSP), and/or any processing component (e.g., a field-programmable gate array (FPGA), an application-specific integrated circuit (ASIC), etc.) that can be programmed perform a function. Memory 606 may include random access memory (RAM), read only memory (ROM), and/or another type of dynamic or static storage device (e.g., flash memory, magnetic memory, optical memory, etc.) that stores information and/or instructions for use by processor 604. In some non-limiting embodiments or aspects, memory 606 may be the same as or similar to ML model management repository 104.

[0146] With continued reference to FIG. 6, storage component 608 may store information and/or software related to the operation and use of device 600. For example, storage component 608 may include a hard disk (e.g., a magnetic disk, an optical disk, a magneto-optic disk, a solid state disk, etc.) and/or another type of computer-readable medium. In some non-limiting embodiments or aspects, storage component 608 may be the same as or similar to ML model management repository 104. Input component 610 may include a component that permits device 600 to receive information, such as via user input (e.g., a touch screen display, a keyboard, a keypad, a mouse, a button, a switch, a microphone, etc.). Additionally or alternatively, input component 610 may include a sensor for sensing information (e.g., a global positioning system (GPS) component, an accelerometer, a gyroscope, an actuator, etc.). Output component 612 may include a component that provides output information from device 600 (e.g., a display, a speaker, one or more light-emitting diodes (LEDs), etc.). Communication interface 614 may include a transceiver-like component (e.g., a transceiver, a separate receiver and transmitter, etc.) that enables device 600 to communicate with other devices, such as via a wired connection, a wireless connection, or a combination of wired and wireless connections. Communication interface 614 may permit device 600 to receive information from another device and/or provide information to another device. For example, communication interface 614 may include an Ethernet interface, an optical interface, a coaxial interface, an infrared interface, a radio frequency (RF) interface, a universal

serial bus (USB) interface, a Wi-Fi® interface, a cellular network interface, and/or the like.

[0147] Device 600 may perform one or more processes described herein. Device 600 may perform these processes based on processor 604 executing software instructions stored by a computer-readable medium, such as memory 606 and/or storage component 608. A computer-readable medium may include any non-transitory memory device. A memory device includes memory space located inside of a single physical storage device or memory space spread across multiple physical storage devices. Software instructions may be read into memory 606 and/or storage component 608 from another computer-readable medium or from another device via communication interface 614. When executed, software instructions stored in memory 606 and/or storage component 608 may cause processor 604 to perform one or more processes described herein. Additionally or alternatively, hardwired circuitry may be used in place of or in combination with software instructions to perform one or more processes described herein. Thus, embodiments described herein are not limited to any specific combination of hardware circuitry and software. The term “configured to,” as used herein, may refer to an arrangement of software, device(s), and/or hardware for performing and/or enabling one or more functions (e.g., actions, processes, steps of a process, and/or the like). For example, “a processor configured to” may refer to a processor that executes software instructions (e.g., program code) that cause the processor to perform one or more functions.

[0148] Although embodiments have been described in detail for the purpose of illustration, it is to be understood that such detail is solely for that purpose and that the disclosure is not limited to the disclosed embodiments or aspects, but, on the contrary, is intended to cover modifications and equivalent arrangements that are within the spirit and scope of the appended claims. For example, it is to be understood that the present disclosure contemplates that, to the extent possible, one or more features of any embodiment or aspect can be combined with one or more features of any other embodiment or aspect.

WHAT IS CLAIMED IS:

1. A computer-implemented method, comprising:
 - receiving, with at least one processor, a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities;
 - generating, with at least one processor, a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein generating the first simulation graph comprises:
 - generating, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models;
 - determining an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities;
 - generating a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and
 - generating second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances;
 - adding, with at least one processor, abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and
 - training, with at least one processor, a machine learning model based on the second simulation graph to provide a trained machine learning model.
2. The computer-implemented method of claim 1, further comprising:
 - generating a graph model based on the plurality of payment transactions involving the plurality of real entities;
 - determining a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities;

wherein generating the plurality of simulation subgraphs of the first simulation graph comprises:

generating the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

3. The computer-implemented method of claim 2, wherein determining the plurality of real subgraphs of the graph model comprises:

selecting one or more nodes to be included in each real subgraph of the plurality of real subgraphs;

clustering the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and

determining a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and

wherein generating the first simulation graph comprises:

generating the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

4. The computer-implemented method of claim 1, wherein adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph comprises:

adding fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and

wherein training the machine learning model comprises:

training a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

5. The computer-implemented method of claim 1, wherein generating the first simulated transaction data for each simulated entity of the plurality of simulated entities comprises:

providing an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity, and wherein the output comprises a prediction of at least one of:

whether a transaction is to be sent by the simulated entity within a time interval,

whether a transaction is to be received by the simulated entity within a time interval, or

whether the simulated entity will be involved in a transaction within a time interval.

6. The computer-implemented method of claim 1, wherein determining the entity type of each simulated entity of the plurality of simulated entities comprises:

generating an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and

clustering each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

7. The computer-implemented method of claim 1, wherein generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models comprises:

generating in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and

generating out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

8. A system, comprising:

at least one processor configured to:

receive a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities;

generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein, when generating the first simulation graph, the at least one processor is configured to:

generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models;

determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities;

generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and

generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances;

add abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and

train a machine learning model based on the second simulation graph to provide a trained machine learning model.

9. The system of claim 8, wherein the at least one processor is further configured to:

generate a graph model based on the plurality of payment transactions involving the plurality of real entities;

determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities; and

wherein, when generating the plurality of simulation subgraphs of the first simulation graph, the at least one processor is configured to:

generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

10. The system of claim 9, wherein, when determining the plurality of real subgraphs of the graph model, the at least one processor is configured to:

select one or more nodes to be included in each real subgraph of the plurality of real subgraphs;

cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and

determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and

wherein, when generating the first simulation graph, the at least one processor is configured to:

generate the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

11. The system of claim 8, wherein, when adding the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, the at least one processor is configured to:

add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and

wherein, when training the machine learning model, the at least one processor is configured to:

train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

12. The system of claim 8, wherein, when generating the first simulated transaction data for each simulated entity of the plurality of simulated entities, the at least one processor is configured to:

provide an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of a real entity, and wherein the output comprises a prediction of at least one of:

whether a transaction is to be sent by the simulated entity within a time interval,

whether a transaction is to be received by the simulated entity within a time interval, or

whether the simulated entity will be involved in a transaction within a time interval.

13. The system of claim 8, wherein, when determining the entity type of each simulated entity of the plurality of simulated entities, the at least one processor is configured to:

generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and

cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

14. The system of claim 8, wherein, when generating the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, the at least one processor is configured to:

generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and

generate out-subgraph simulated transaction data for transactions involving entities in two different simulation subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.

15. A computer program product comprising at least one non-transitory computer-readable medium including one or more instructions that, when executed by at least one processor, cause the at least one processor to:

receive a dataset comprising a plurality of data instances, wherein the plurality of data instances is associated with a plurality of payment transactions involving a plurality of real entities;

generate a first simulation graph representing a plurality of simulated payment transactions involving a plurality of simulated entities based on the plurality of data instances, wherein, the one or more instructions that cause the at least one processor to generate the first simulation graph, cause the at least one processor to:

generate, based on the plurality of data instances, first simulated transaction data for each simulated entity of the plurality of simulated entities using one or more first transformer machine learning models;

determine an entity type of each simulated entity of the plurality of simulated entities based on the first simulated transaction data for each simulated entity of the plurality of simulated entities;

generate a plurality of simulation subgraphs of the first simulation graph based on the entity type of each simulated entity of the plurality of simulated entities; and

generate second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models based on the plurality of data instances;

add abnormal transaction patterns to the first simulation graph to provide a second simulation graph; and

train a machine learning model based on the second simulation graph to provide a trained machine learning model.

16. The computer program product of claim 15, wherein the one or more instructions further cause the at least one processor to:

generate a graph model based on the plurality of payment transactions involving the plurality of real entities;

determine a plurality of real subgraphs of the graph model based on transaction behavior of each real entity of the plurality of real entities; and

wherein, the one or more instructions that cause the at least one processor to generate the plurality of simulation subgraphs of the first simulation graph, cause the at least one processor to:

generate the plurality of simulation subgraphs of the first simulation graph based on the plurality of real subgraphs of the graph model.

17. The computer program product of claim 16, wherein, the one or more instructions that cause the at least one processor to determine the plurality of real subgraphs of the graph model, cause the at least one processor to:

select one or more nodes to be included in each real subgraph of the plurality of real subgraphs;

cluster the plurality of real subgraphs into a plurality of types of real subgraph clusters based on a measure of similarity between the plurality of types of real subgraphs to provide a plurality of real subgraph clusters; and

determine a ratio of each type of real subgraph clusters to all types of real subgraph clusters; and

wherein, when generating the first simulation graph, the at least one processor is configured to:

generate the first simulation graph to include a plurality of types of simulation subgraphs, wherein a ratio of each type of simulated subgraph cluster to all types of simulated subgraph clusters is the same as a ratio of each type of real subgraph cluster to all types of real subgraph clusters.

18. The computer program product of claim 15, wherein, the one or more instructions that cause the at least one processor to add the abnormal transaction patterns to the first simulation graph to provide the second simulation graph, cause the at least one processor to:

add fraudulent transaction patterns to the first simulation graph to provide the second simulation graph; and

wherein, the one or more instructions that cause the at least one processor to train the machine learning model, cause the at least one processor to:

train a fraud classification machine learning model based on the second simulation graph to provide a trained fraud classification machine learning model.

19. The computer program product of claim 15, wherein, the one or more instructions that cause the at least one processor to generate the first simulated transaction data for each simulated entity of the plurality of simulated entities, cause the at least one processor to:

provide an input to the one or more first transformer machine learning models to generate an output based on the input, wherein the input comprises a plurality of time related transaction features of a transaction involving a real entity corresponding to the simulated entity and a plurality of aggregated transaction features associated with a last transaction of the real entity, and wherein the output comprises a prediction of at least one of:

whether a transaction is to be sent by the simulated entity within a time interval,

whether a transaction is to be received by the simulated entity within a time interval, or

whether the simulated entity will be involved in a transaction within a time interval.

20. The computer program product of claim 15, wherein, the one or more instructions that cause the at least one processor to determine the entity type of each simulated entity of the plurality of simulated entities, cause the at least one processor to:

generate an embedding for each simulated entity of the plurality of simulated entities using the one or more first transformer machine learning models; and

cluster each simulated entity of the plurality of simulated entities using a K-means clustering algorithm based on a plurality of embeddings for the plurality of simulated entities to provide a plurality of clusters, wherein each cluster of the plurality of clusters is associated with an entity type of a plurality of entity types.

21. The computer program product of claim 15, wherein, the one or more instructions that cause the at least one processor to generate the second simulated transaction data for each simulated entity of the plurality of simulated entities using one or more second transformer machine learning models, cause the at least one processor to:

generate in-subgraph simulated transaction data for transactions involving entities within each simulation subgraph of the plurality of simulation subgraphs using an in-subgraph transformer machine learning model; and

generate out-subgraph simulated transaction data for transactions involving entities in two different subgraphs of the plurality of simulation subgraphs using an out-subgraph transformer machine learning model.