Research Internship Report — Process Mining using Egocentric datasets

Research Internship Report

Pooja Gowda, s4410963, p.gowda@student.rug.nl
Supervisors: Prof. Dr. Dimka Karastoyanova & Asst.Prof. Estefanía Talavera Martínez, PHD

Abstract: Process Mining techniques involve creating process models from the set or pattern of activities carried out in a process. Now, we suppose every individual has their routine, which is a pattern of activities and by observing these patterns we can infer the lifestyle of a person we can even find many insights regarding his/her behavior. By doing so we can identify many hidden mental health issues, working environment issues, and so on. These activities' data(audio/video/images) specific to one person's daily regular activities have been logged which is termed here as ego-centric datasets. In this research work, we will explore these datasets collected from different sources of the world and apply process mining algorithms using the Prom plugin. By doing so, we intend to discover the potential of involving process mining techniques with data mining to explore significant research directions.

Contents

1 Introduction

In the past few years, the involvement of data science and data mining in building process models has arisen. By using the data mining techniques the possibility to discover process models directly by using data is higher. This possibility has provided an area to explore and research the potential directions. The goal of this research is to identify ways to interpret the ego centric data while using it as an event log from which we can derive process models and finally explore research direction.

First, I began by researching the prospects of applying process mining techniques to produce the process model using ego centric data set. The section 2 intends to provide some background knowledge regarding the data set type used, and in the section 3 we will discuss few previous related literature work to give some brief idea about the research work.

In The further section we will discuss about process mining algorithms,methods , miners , Prom plugin briefly. This information will be very significant while understanding implementation and results that is obtained.

2 Background

In this section, we will discuss some background knowledge used in this research work. This will help the readers to understand the concepts that will be discussed throughout this research work. A process model is a collection of events from start to end or a process. By using ego centric data set we can create a process model which will help us to visualize how one person's normal behavior by looking at the events conducted on daily basis. While attempting to do so we can end producing the process models with higher or lower detailing with the respect to events, we might have to filter out certain events. We will discuss the basic concepts that are used in order to understand the converting procedure of egocentric data to process models and process mining can be applied to discover automatically process models.

2.1 Egocentric Data sets

The data set that is used for the research purpose is been downloaded from https://ego4d-data.org, which is illustrated using Figure 2.1. The data set is diverse in its geographic range, scenarios, participants, and captured modalities (7).Using various off-the-shelf head-mounted cameras: GoPro, Vuzix Blade, Pupil Labs, ZShades, OR- DRO EP6, iVue Rincon 1080, and Weeview. The egocentric data set is a data set captured from the first-person or "egocentric" point of view where we capture the activities with the respect to the eyes of a user directly (7).

2.2 Event Log

To create or discover a process model an event log plays a significant role. Each event log is formed by a set of traces that gives insight into a set of events, activities, or scenarios that occurred to complete a process in one person's daily life. Every event must consist of attributes that describe the type of activity carried out. Timestamp can be added to understand in detail the activities carried in particular timeline of a day. The figure2.2

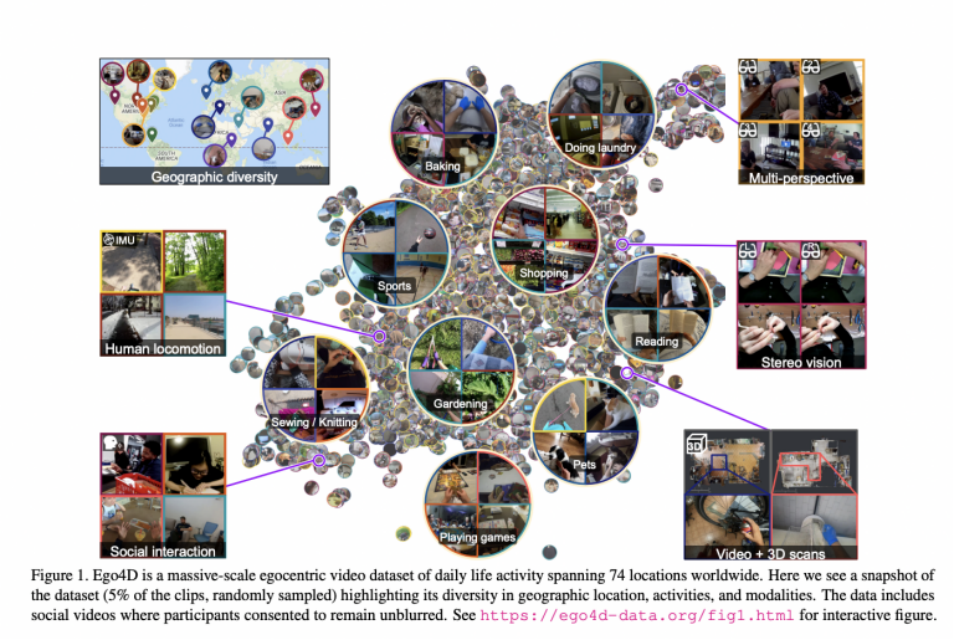


Figure 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

Figure 2.1: Egocentric data sample collected from ego4d-data.org(1)

is an example of event logs which can be used. The event log can be identified only by

Patient ID	Event ID	Activity	Start Date	Department
1045	12	Admission to hospital	3 January 2017 09:47	Cardiovascular surgery
1045	13	Surgery started	4 January 2017 08:00	Cardiovascular surgery
1045	14	Transfer to ICU	4 January 2017 14:53	Cardiovascular surgery
1045	15	Surgery finished	4 January 2017 14:55	Cardiovascular surgery
1045	16	Surgery started	5 January 2017 06:15	Cardiovascular surgery
1045	17	Surgery finished	5 January 2017 07:50	Cardiovascular surgery
1045	18	Transfer to service	13 January 2017 11:20	Cardiovascular surgery
1045	19	Discharged	20 January 2017 11:49	Cardiovascular surgery

Figure 2.2: Event Log Sample (6)

the attribute it carries, as mentioned previously it should describe a particular activity or event occurrence. In our scenario, 'scenarios' specifies the event carried out in a particular time period. Hence, we consider 'scenarios' as the event log in this experiment.

2.3 Notations

The Transitions and the mathematical sets are represented with the notations to describe a sequence of events. So we will be discussing few examples and notations below.

2.3.1 Petri Nets

Petri nets are one of the most reasonable studied process modeling languages, providing a space for concurrency modeling. We can execute the Petri nets and even analyze them. A Petri Net includes transitions and placements that generate a static network. This is managed using the firing rule, tokens are used to simulate the information flow.
Definition 1. A Petri Net is a triplet N=(P,T,F) where P is a finite set of places, T is a finite set of transitions such that $P \cap T = \varnothing$, and $F \subseteq (P \times T) \cup (T \times P)$ is a set of directed arcs,called the *flow relation.*A marked Petri net is a pair (N,M), where N =

(P,T,F) is a Petri net and where M ∈ B(P) is a multi-set over P denoting the marking of the net. The set of all marked Petri nets is denoted N(5).

2.3.2 Transition Systems

The transition system is one of the primary notations used for process modeling. The system is a combination of states and transitions with initial and final states labeled uniquely. Transitions are represented using arcs; sometimes more than one can have the same label(5).
Definition 2. A transition is a triplet TS= (S,A,T) where S is the set of the states, A is set of activities, and T ⊆ S X A X S is the set of transitions. $S^{start} \subseteq$ S is the set of initial states or start states and $S^{end} \subseteq$S is the set of final states or known as accept states(5).

2.3.3 Business Process Model and Notation

The Business Process Modeling Notation (BPMN) is one of the most used languages in modeling business processes. It is supported by a tool and is standardized by the OMG. Also, BPMN supports nested workflows which come in use while representing sub-processes (5).

2.4 Process Mining

Process mining acts like a missing link between data science and process science. The process mining starts from event data and uses process models in various ways. Process mining is process-centric and as well as data-driven, by using process models and event data a range of conformance and performance-based analysis can be done(5).

2.5 Prom Plugin

The functionalities carried out in businesses or any organization consists of several processes, which are significant to observe the activities and optimize the overall activities carried out. There are several ways to monitor and analyze these processes, prom plugin provides a process mining inclusive tool with several algorithms and miners.

The extraction of significant information concerned with processes observed from its event logs is process mining all about. The insights are obtained into perspectives, namely control flow, performance, data, and organization. Process mining techniques are in the form of plug-ins which are supported by the ProM tool, as seen in Figure 2.3. This is an extensible framework that is platform-independent and is available as an open source for free of cost.

3 Previous Related Works

In 2017, the author discusses process mining techniques that concentrate on extracting insights from the event logs. They shed light on the scenario where the event log is too fine-grained with many low-level event logs. Hence, the author proposes to discover the structure in the process by first abstracting the event log to a higher level of granularity.

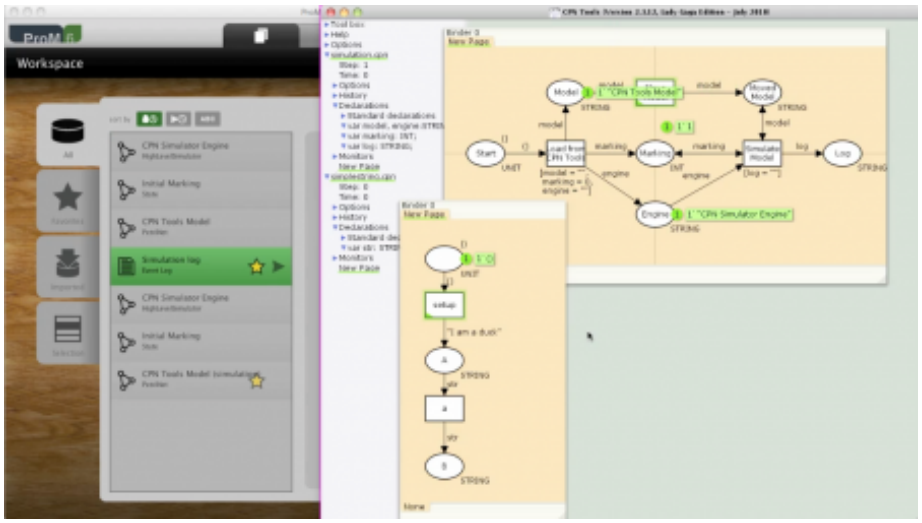


Figure 2.3: ProM Framework (3)

A method is discussed which generates feature vector representations of events established on XES extensions and an approach known as Conditional Random Fields is used. The approach steps go as follows:

- A set of annotated traces, which are traces where the high-level event belongs is known for all low-level events in the trace.
- It is a set of unannotated traces, which are traces where the low-level events are not mapped to the high-level events.

The conditional Random Fields are trained on annotated traces to create a probabilistic mapping from the low level to the high level of events. After obtaining these mappings, it is applied to the unannotated traces in order to assess the respective high level event for each low level event. Checked few case studies and implementation of this approach.

While in 2020, The author introduces The topic Modelling(TM) technique, which is used to observe the correlated elements of the individual's day. The author uses TM as an unsupervised approach for the analysis of behavioral traits to detect routines from egocentric to understand the patterns observed in the user's daily behavior. They even discuss a proposed model which characterizes the egocentric photo stream data to be classified as Routine and Non Routine related days. The Data is a set of user files with egocentric images of each user logged at different timestamps.

In 2021, As part of the bachelor thesis the author worked on exploring process discovery using Egocentric photo streams. The photo streams are collected by different users on different dates, by exploring every user's daily activity using the ProM framework the author presented some insightful results.

4 Mining Methods/Algorithms

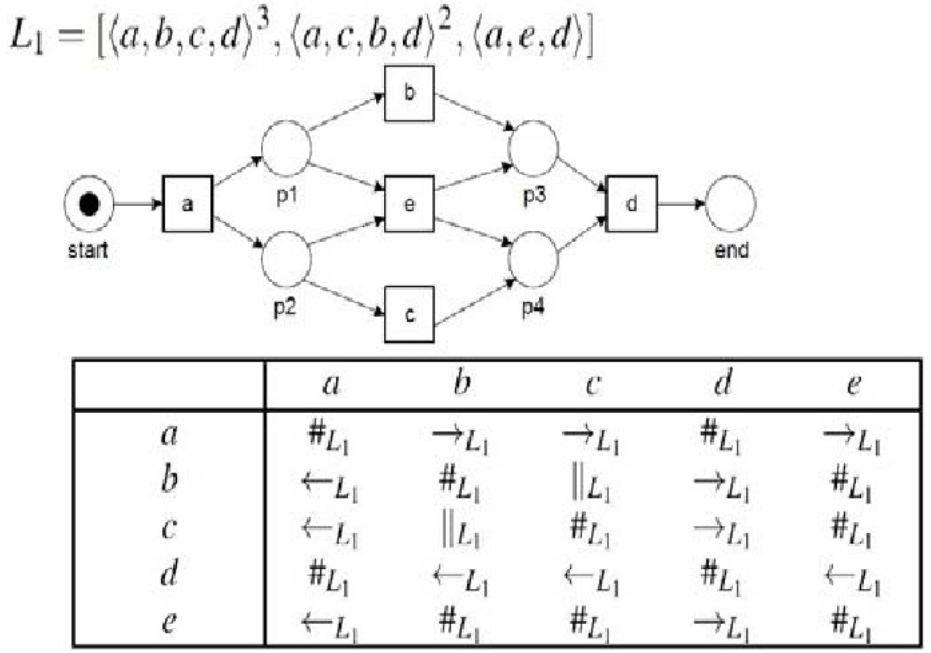
The ego centric data is used in process mining techniques using several algorithms. In this section we will discuss about various algorithms used while experimenting with the datasets.

4.1 Alpha Miner

Alpha Miner is the algorithm that helps in filling the link between event logs and the process model. This algorithm helps in building process models based on event logs with relations and causalities between the steps of processes.

The Alpha miner generates:

- A Petri net model with all the transitions being visible, unique, and correspond to the classified events.
- The initial marking — Describes the status of the Petri net model when the execution starts.
- The final marking — Describes the status of the Petri net model when the execution ends(2).



	a	b	c	d	e
a	$\#_{L_1}$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$
b	$\leftarrow_{L_1}$	$\#_{L_1}$	$\|_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
c	$\leftarrow_{L_1}$	$\|_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
d	$\#_{L_1}$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\leftarrow_{L_1}$
e	$\leftarrow_{L_1}$	$\#_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$

Figure 4.1: Model obtained by applying Alpha Miner (4)

4.2 Inductive Petri Net Miner

The Splits between the start and final state of the process in a log are detected by using Inductive Miner algorithm. Types of splits identified are as follows (2):

- Sequential
- Parallel
- Concurrent
- Loop

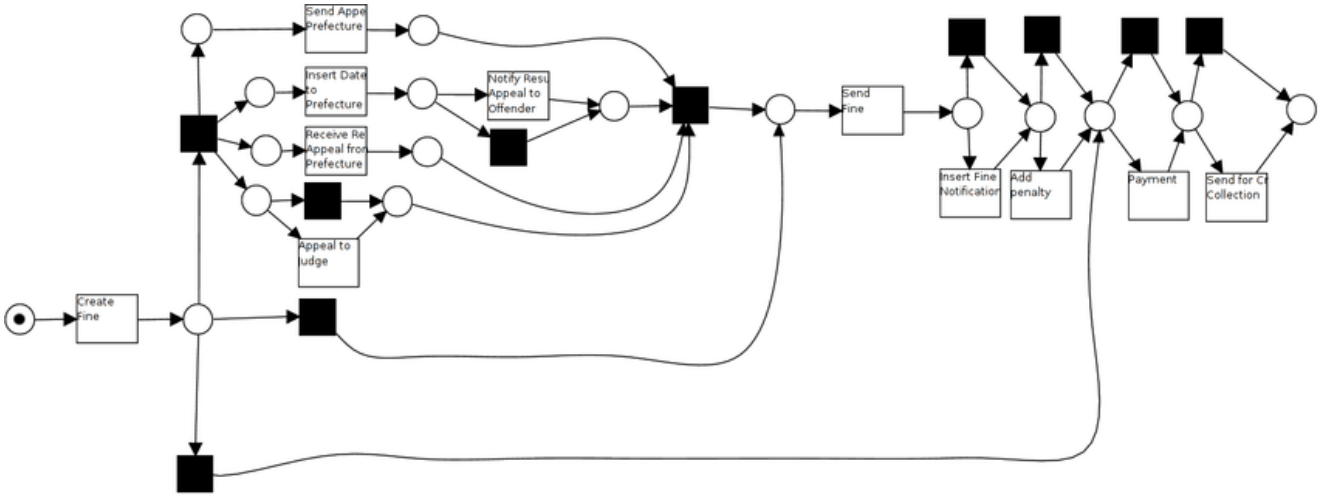


Figure 4.2: Model obtained by applying Inductive Miner(8)

On identifying the splits, this repeats until a base case is found. The models generated utilizes hidden transitions for loop splits. By using this algorithm infrequent behaviors are dropped. Hence, providing a model which is favorably fitting and behaviorally valid

4.3 Evolutionary Tree Miner (ETMd) Process Tree Miner

ETmd is a genetic algorithm that conducts the discovery process depending on the four specific quality dimensions from the discovered model which are (2):

1. Fitness
2. Precision
3. Generalization
4. Complexity

4.4 Discover Graph (Causal Net Mining)

The causal relations extracted from the event log and the background knowledge on constraints over the topology of the to be produced models are used.

5 Implementation

5.1 Preprocessing

Most of the algorithms require data with a limited amount of changes, So I have kept the data as much as possible unchanged. Since the data is extracted from a different video source, the data was split into two several CSV files concerning the video source with the specific case number to be 1,2, or 3(representing process occurring in different places). further, I even merged these CSV files into one to execute miners on this data. You can find code in Appendix.‘

5.2 Converting to XES

The parsed data is stored in CSV with case identifiers. As per XES format having an Id and the case number is important for running the data in Prom. Prom provides a plugin that converts CSV to XES format, using this I get 'XMLEventlog'. Further, I applied various process mining algorithms regarding which I will discuss the implementation in the next section.

5.3 Apply Process Mining Techniques

Due to the difficulty in running the mining algorithms several times on the dataset by varying the parameters every time we run, I opted to work on smaller subsets of data. We applied mining algorithms to this dataset using the ProM tool.

6 Results

6.1 Inductive Petri Net Miner

The Process Mining began with the Inductive Petri Net Miner(infrequent(IMF)) algorithm. This was one of the successful miners. While experimenting it was observed as you reduce noise from 0.20 to 0.10 the self-loops increases. The generated models on several cases is in Appendix.

6.2 ETMd Process Tree Miner

On application of this algorithm the following variables we asked to tinker: precision,replay fitness, generalization and simplicity. On implementing this algorithm it takes longer run time to mine, so I didn't change parameter values. The resulted process model can be found in github repository `https://github.com/poojagowda7818/Research_Internship` as the image width seem to not fit.

7 Conclusions

The goal of this research internship was to research the different available methods by applying process mining where the data is egocentric. In application, I had to explore and propose potential directions.

One of the biggest challenges in this project was with the dataset. It was a bunch of many datasets with no particular timestamps, hence it took a long time to explore the dataset. The dataset used is collected from a different region of the world with logged data of daily activities.

With the data set downloaded the initial analysis as per the implementation and results section shows a positive sign towards the potential of including process mining techniques over generally logged data. I applied on Two different miners while testing I determined that the inductive miner gave better results comparatively, as it discovered the path of the activity. Since the logged data is a set of activities detected in a video sequentially occurring in a day. Using ETmd it provided a Petri net with a very vague representation, which was hard to comprehend.

We can conclude from this experimentation there is potentiality in using ego-centric data sets to generate process models. Also, by going further in this research direction, we can discover what the number of occurrences of a certain event tells about the health status of the user and set of people living in that area.

8 Future Work

The Future work intends to work extensively in analysing the process model creating several research directions:

- To detect and inform the unusual changes in the daily activities logged and improve lifestyle of the user by suggesting activities to improve the routine.
- To create daily schedules as per the daily activities.

There was plenty of ways to improve the results. The list is as follows:

1. **Data sets**: with timestamp(dd-mm-yyyy) would help to understand when did the activity start and end.
2. **Using a several algorithm** Applying the several algorithms on the data set would definitely turn out to be a value added to the results.

References

[1] Egocentric = https://venturebeat.com/2021/10/14/facebook-introduces-dataset-and-benchmarks-to-make-ai-more-egocentric/, note = Accessed: 19-05-2022.

[2] Process mining = https://research.aimultiple.com/process-mining-algorithms, note = Accessed: 09-05-2022.

[3] Prom tool = https://westergaard.eu/2010/07/cosimulating-cpn-models-and-prom-plugins-and-application-to-prom-orchestration/, note = Accessed: 09-05-2022.

[4] Process mining using -algorithm as a tool (a case study of student registration? *Computer Science 2012 Tenth International Conference on ICT and Knowledge Engineering*, 2012.

[5] Tugba Gurgen Erdogan and Ayca Tarhan. Process mining - data science in action. 2016.

[6] Tugba Gurgen Erdogan and Ayca Tarhan. A goal-driven evaluation method based on process mining for healthcare processes. *MDPI*, 2018.

[7] Eugene Byrne Zachary Chavis Antonino Furnari Rohit Girdhar Jackson Hamburger Kristen Grauman, Andrew Westbury. Ego4d: Around the world in 3,000 hours of egocentric video. *Computer Vision and Pattern Recognition*, 1, 2021.

[8] Sebastiaan J. van Zelst Wil M.P. van der Aalst, Alfredo Bolt. Rapidprom: Mine your processes and not just your data. 2017.

A Appendix

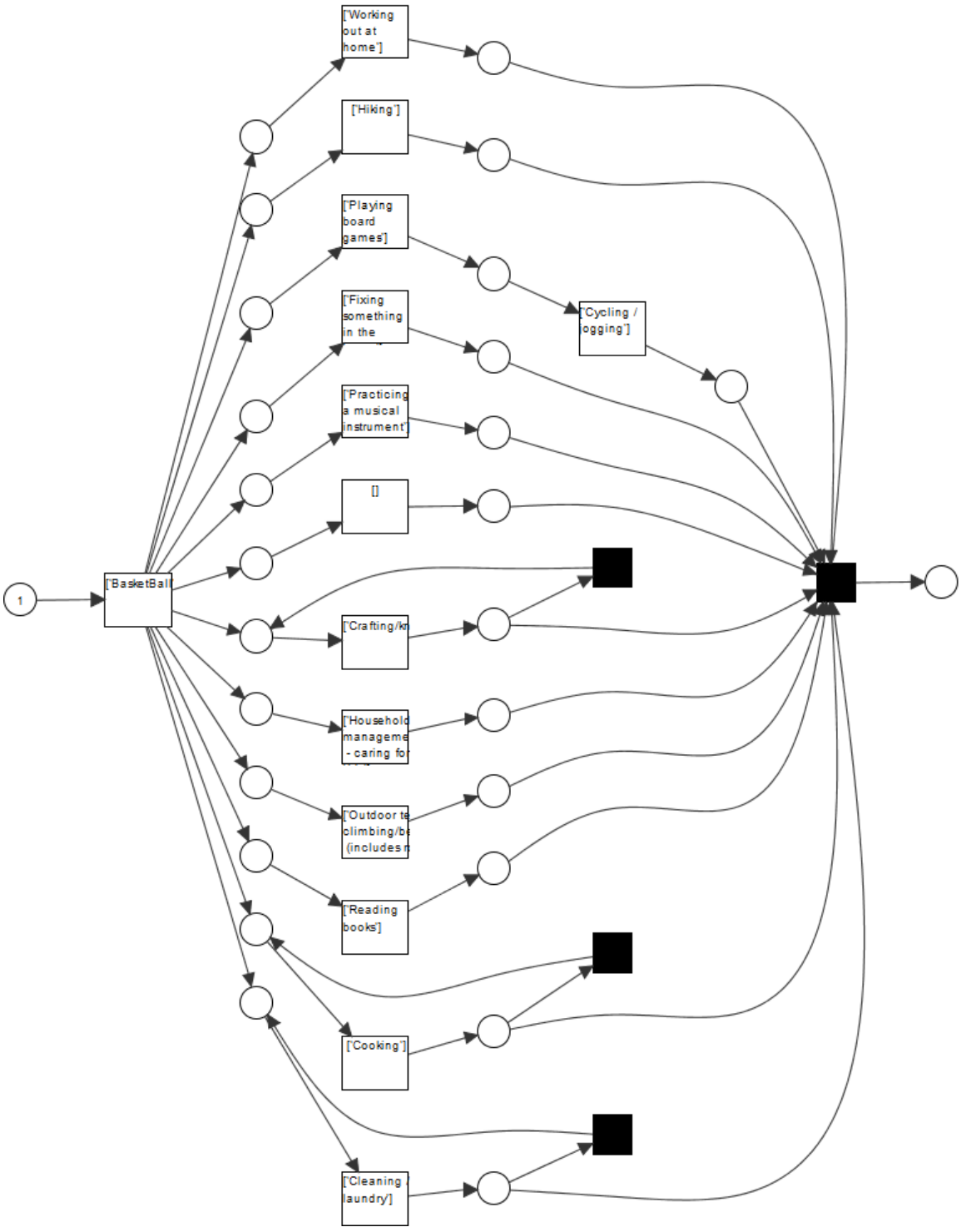


Figure A.1: The resultant process model using Inductive miner on data logged from Bristol location(case3)

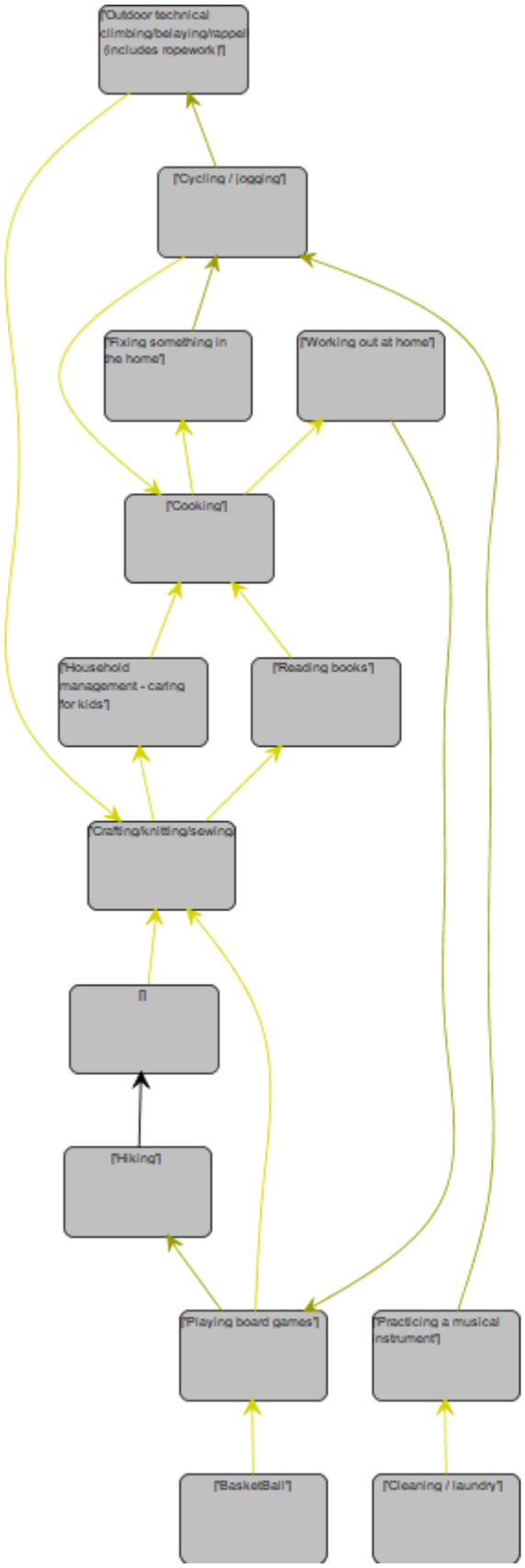


Figure A.2: The resultant process model using discovery graph on data logged from Bristol location case(3)

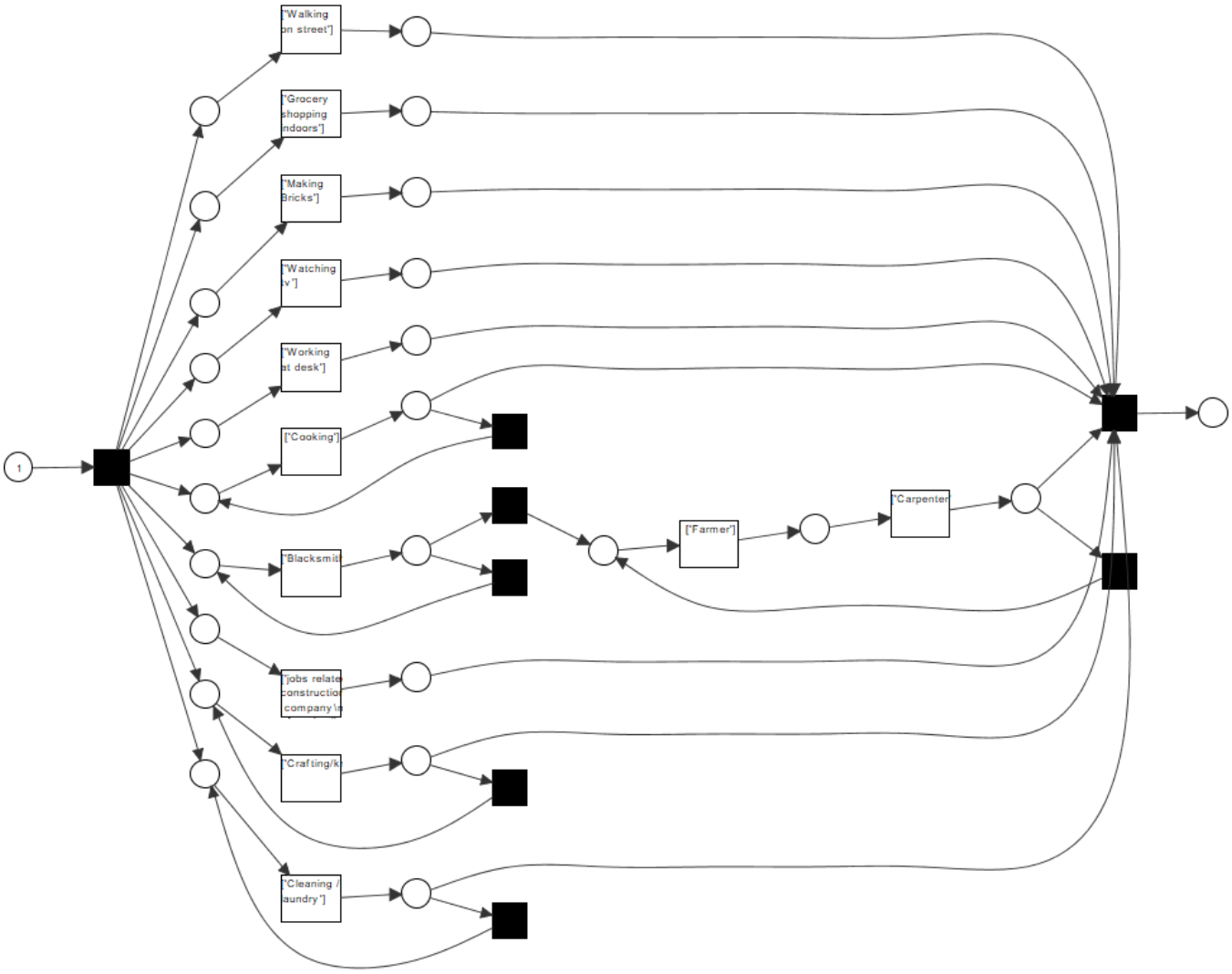


Figure A.3: The resultant process model using Inductive miner on data logged from iiith location(case2)

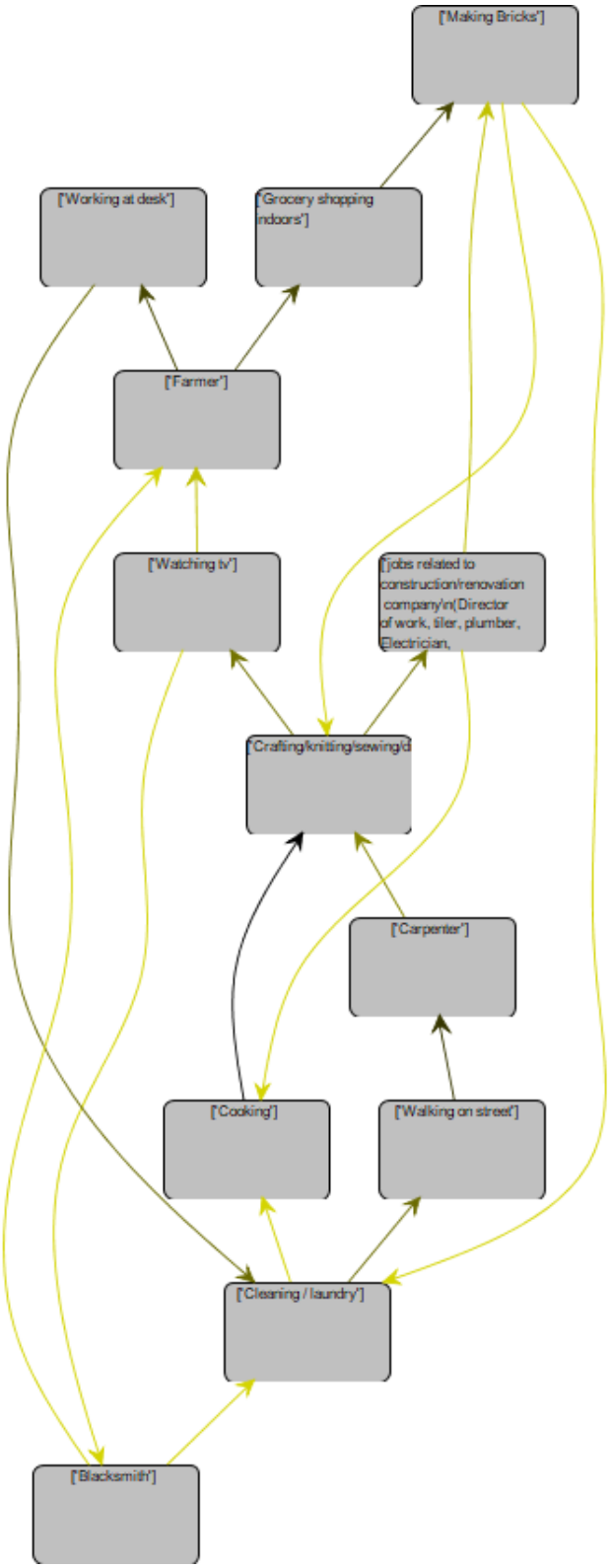


Figure A.4: The resultant process model using discovery graph on data logged from iiith location case(2)

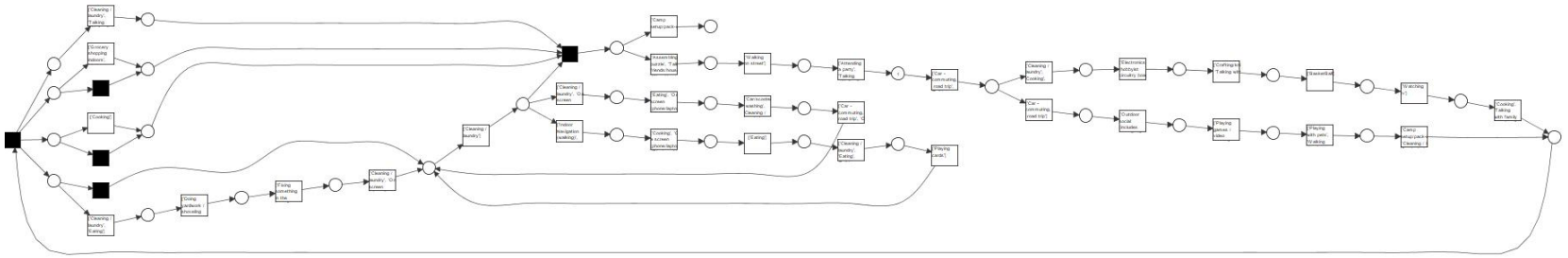


Figure A.5: The resultant process model using inductive miner on data logged from minnesota location case(1) with noise 0.20

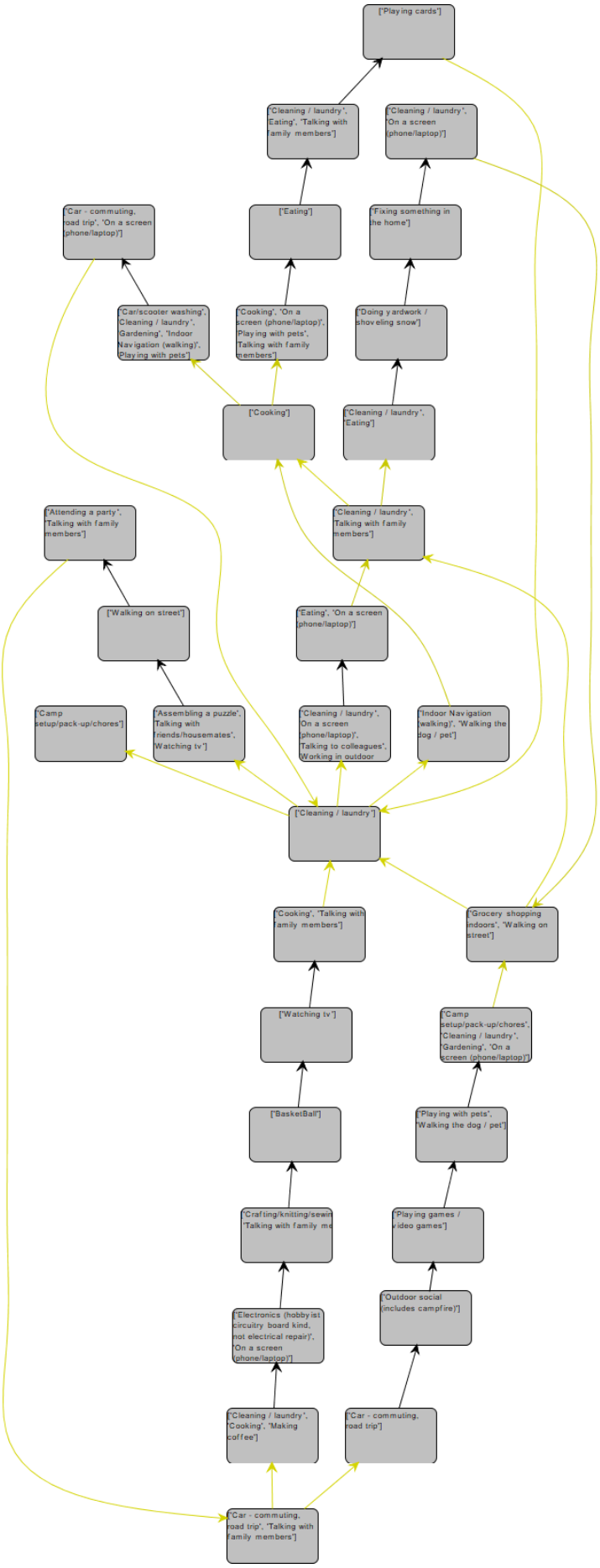


Figure A.6: The resultant process model using discovery graph on data logged from minnesota location case(1)

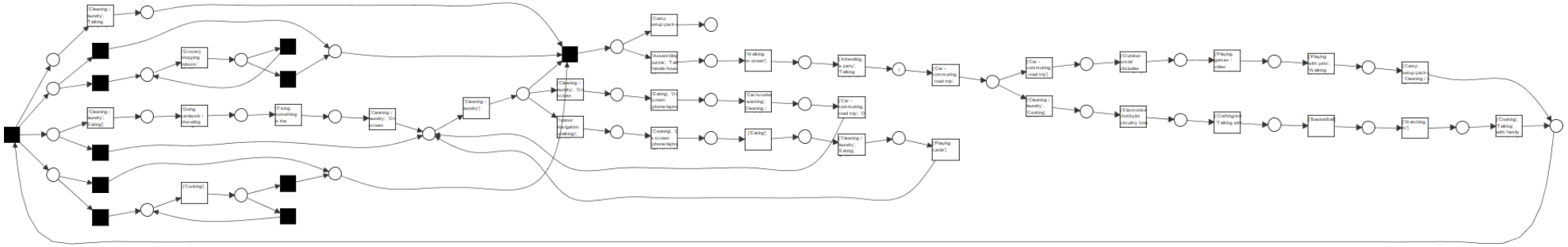


Figure A.7: The resultant process model using inductive miner on data logged from minnesota location case(1) with noise 0.10

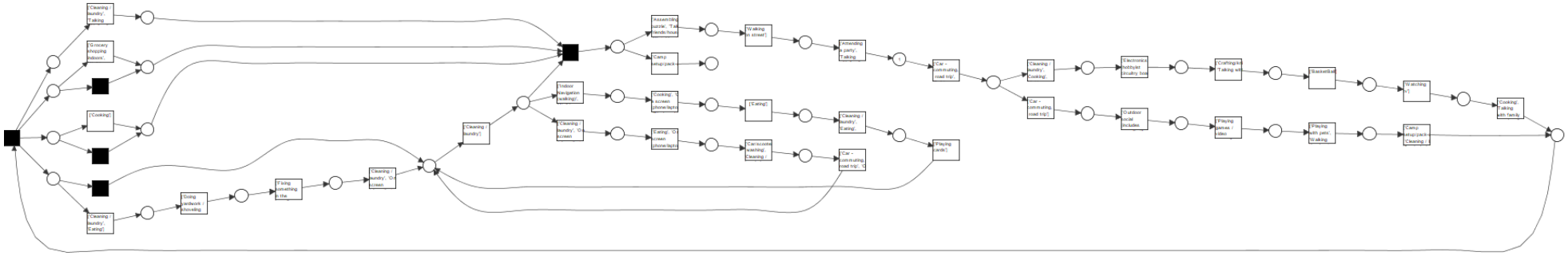


Figure A.8: The resultant process model using inductive miner on data logged from minnesota location case(1) with noise 0.30