Gas Detection and Source Localization: A Bayesian Approach

Gregor Pavlin
D-CIS Lab
Thales Research and Technology
Delft, The Netherlands
& Informatics Institute
University of Amsterdam
Email: gregor.pavlin@d-cis.nl

Patrick de Oude
D-CIS Lab
Thales Research and Technology
Delft, The Netherlands
& Informatics Institute
University of Amsterdam
Email: patrick.deoude@d-cis.nl

Franck Mignet
D-CIS Lab
Thales Research and Technology
Delft, The Netherlands
Email: franck.mignet@d-cis.nl

Abstract—This paper discusses modeling solutions that support detection of gaseous chemical substances and source localization in applications that are characterized by large numbers of noisy information sources, absence of calibrated concentration measurements and lack of detailed knowledge about the physical processes. In particular, we introduce a solution based on discrete Bayesian networks which allows tractable exploitation of large quantities of spatio-temporally distributed heterogeneous observations. The emphasis is on using coarse models avoiding assumptions about detailed aspects of the gas propagation processes. By considering properties of Bayesian networks we discuss the consequences of modeling simplifications and show with the help of simulations that the resulting inference processes are robust with respect to the modeling deviations.

Keywords: Bayesian inference, Gas detection, Leak localization.

I. INTRODUCTION

This paper discusses a tractable solution to a relevant class of contemporary environmental monitoring problems. In particular, the focus is on detection of harmful gaseous substances and localization of sources in densely populated industrial areas with high concentration of chemical industrial facilities. An example of such a place is the larger Port of Rotterdam area hosting large petrochemical terminals and plants. The risk of exposure to harmful and annoying gaseous pollutants released during production or through accidents is in such an environment high. The capability to quickly detect such gaseous substances and to locate the sources is critical for the effective mitigation of the impacts of chemical incidents and industrial pollution.

However, approaches using absolute concentration measurements, such as for example [3], [4], [10], are not suitable for the targeted settings. Namely, adequate networks of calibrated chemical sensors measuring concentrations are economically not feasible; for each gas a specific type of sensors is required, calibration is very difficult in outdoor settings and, consequently, calibrated sensors are expensive. In addition, fine grained models accurately describing the physical aspects of gas dispersion in terms of concentration distributions are usually not available. Thus, we must rely on coarse dispersion models and anomaly detectors using signals from uncalibrated, noisy chemical sensors.

The key to real world solutions are domain models that adequately capture complex relations between large quantities of spatially and temporally distributed observations and hidden phenomena. It turns out that probabilistic causal models support efficient and robust solutions, since the observations can often be viewed as outcomes of causal stochastic processes. By understanding the underlying causal mechanisms and by considering the observations, the hidden phenomena can be inferred. Such causal processes can efficiently be modeled with the help of Bayesian networks (BN) [2], [9].

Due to the inherent complexity of real world processes, the resulting models are characterized by many variables and relations corresponding to large numbers of parameters. The estimation of those modeling parameters requires large quantities of data and rich domain expertise. However, the available quantities of real world data and the expertise are often limited and do not support creation of perfect models. In case of BNs, inevitable modeling deviations occur in parameters as well as in the structure. In other words, the domain models themselves are associated with significant uncertainties regarding the parameters and relations. The main question is whether we can build reliable detection and localization systems, despite such modeling imperfections.

By taking into account the properties of BNs, we derived a simplified model which supports efficient processing and tractable construction of domain models without compromising the performance of the overall detection and localization processes. The impact of the modeling simplifications is investigated by considering the modeling and inference principles of BNs in controlled experiments by systematically introducing modeling deviations. The presented solution has been applied in a prototype system which is currently being evaluated in a real world setting using uncalibrated sensors and human reports.

The paper is organized as follows: in section II the targeted detection and leak localization problem is introduced, section III discusses the modeling principles, section IV provides an overview of the usage of the resulting Bayesian models and
section V discusses a subset of experiments in which the solution was evaluated with the help of simulated data.

II. THE LEAK LOCALIZATION CHALLENGE

In this paper we assume plumes caused by gas leaks which are active for a significant period of time. In such cases the plume life-cycle can be subdivided into three phases. In the first phase the plume spreads, i.e. the area in which the gas is detectable by the sensors and people increases. In this phase the concentrations at different locations in the area of influence steadily increase. During the second phase gas concentrations at any point in the affected area remain above the detection threshold. In addition, if the concentration at a particular location exceeded a certain detection threshold during the first phase, then the concentration will remain above this threshold throughout the second phase. The third phase takes place after the source is eliminated and the gas concentrations are gradually reduced. In this work we focus on the first two phases, which are critical for the timely detection of gases and coarse localization of sources.

The main purpose of the presented approach is to detect the presence of the gas at different locations within the zone affected by the plume and to estimate the source of the pollution as quickly as possible. This must be achieved by using a network of cheap chemical sensors, which are not calibrated. In combination with suitable algorithms, such sensors can be used for simple detection of anomalies, i.e. unusual gas mixtures caused by industrial pollution.

As the system receives a trigger, a positive detection of a specific gas or an industrial anomaly, potential pollution sources are determined by taking into account the wind speed, wind direction and simple gas dispersion models. In the presented approach, for each possible source s we create a cone, a triangular area whose axis is parallel to the wind and its vertex is defined at the potential source (see example in figure 1). If the triggering sensor is within the cone associated with this potential source, the system creates a hypothesis h that s is indeed the source. Note that the cone is an extremely crude gas dispersion model which represents an area in which gas concentrations are likely to exceed the levels required for the detection of gases/anomalies. The model is chosen such that it covers an area which is significantly greater than the actual area in which such concentrations would be observed.

Moreover, the system ranks the hypotheses by computing posterior probability $P(h|e_i)$ for each hypothesis h, where $e_i$ denotes all observations that can be associated with h. The computation of $P(h|e_i)$ is based on a simple domain model which describes gas propagation and typical observation sequences (i.e. patterns), given hypothesis h were true.

III. MODELS OF OBSERVATION GENERATING PROCESSES

The ranking of the hypotheses is based on domain models which capture spatio-temporal relations between the hidden states of interest and observations dispersed throughout a large area. Such models provide a mapping between the hidden states of interest and observations (i.e. patterns). Moreover, we assume discrete domain models, which facilitate construction of fusion systems and efficient reasoning. The models are based on a crude discretization of the areas influenced by the plume, i.e. regions within which critical gas concentrations are exceeded. An example of such a discretization is shown in figure 1. In section IV-B we discuss the impact of such a discretization on the overall performance of the system. In addition, temporal aspects of the dynamic processes are captured with the help of modeling time slices, each representing a snapshot of the states at a specific time step.

A key point in the presented approach is that observations obtained in the aforementioned monitoring processes can be viewed as outcomes of stochastic causal processes, where hidden events cause observations according to certain probability distributions. Therefore, we make use of causal Bayesian networks (BN), which support efficient construction of discrete probabilistic models of stochastic processes generating observation.

The causal model shown in figure 2 captures correlations between heterogeneous, spatially and temporally distributed observations and hidden states. The presence of gas in the domain at time $t(j)$ is represented with the help of a set of binary variables $G_1^{(j)}, G_2^{(j)}, \ldots, G_N^{(j)}$. States of each individual binary variable $G_i^{(j)}$ correspond to the situations in which the gas concentration at location/segment $i$ at time $t(j)$ was above or below some critical level, respectively. More precisely, $G_i^{(j)} = true$ if the plume front has reached the boundary between segments $i$ and $i+1$. Moreover, nodes $G_E^{(j)}$ represent the state of the immediate neighborhood $l_E$ of segment $i$ up wind from the leak. Nodes $L_i^{(j)}$ represent the hypothesis about the leak. State $L_i^{(j)} = true$ corresponds to the hypothesis that the leak is present at $l_i$ at time $t(j)$.

1In this paper we assume that location $l_i$ corresponds to a segment in the crude plume profile shown in figure 1.
The arrows between the state nodes in this model represent direct dependencies between the variables. The BN shown in figure 2 captures relations which correspond to a Left to Right Hidden Markov Model [1]. Namely, the propagation is assumed to have a distinctive direction; the gas can propagate from location $l_1$ to $l_2$, from $l_2$ to $l_3$ and so on.

A. Conditional Probability Tables

The strength of causal relations represented by the process model are captured by the conditional probability tables (CPT). The parameters defining those tables are chosen, such that the main characteristics of the modeled physical processes are captured. We first discuss the choice of parameters in the CPT $P(G_{i}^{(j)}|G_{i}^{(j-1)}, G_{i-1}^{(j-1)})$ relating hidden states of variables capturing plume propagation shown in table I.

Since we are dealing with plumes, the concentrations at location $l_i$ remain high for a significant period of time after the gas reached $l_i$. Thus, the states at such locations can be considered static for the duration of the estimation process. This is reflected by the deterministic transitions between the states of variables representing the presence of gas at the same location at different times; i.e. $P(G_{i}^{(j)} = true|G_{i}^{(j-1)} = true, G_{i-1}^{(j-1)} = true) = 1$.

In addition, if the gas was not present at $l_i$ at time $t(j-1)$, the state of $G_{i}^{(j)}$ is directly influenced by the state at a neighboring segment at time $t(j-1)$, represented by variable $G_{i-1}^{(j-1)}$. This influence is captured by the conditional probability $P(G_{i}^{(j)} = true|G_{i}^{(j-1)} = false, G_{i-1}^{(j-1)} = true) = p > 0.5$.

There is also a small chance that a new source materializes at any location $l_i$ at any time step $t(j)$. This means that $G_{i}^{(j)} = true$ even though $G_{i}^{(j-1)} = false$ and $G_{i-1}^{(j-1)} = false$. Since the prior probability of such an event at any location is typically very small, the probability $P(G_{i}^{(j)} = true|G_{i}^{(j-1)} = false, G_{i-1}^{(j-1)} = false) = 1 - q$ is small as well. In other words, $q$ is close to 1.

Moreover, if the gas was not present at location $l_{i-1}$ at time $t(j-1)$, i.e. $G_{i}^{(j-1)} = false$, while at segment $l_i$, the gas was present at time $t(j-1)$, i.e. $G_{i}^{(j-1)} = true$, then we are not dealing with a plume, but with a traveling cloud. Therefore, in the next time slice corresponding to $t(j)$, the cloud is likely to clear segment $l_i$ as well. In other words, the chance of having no gas at location $l_i$ at time $t(j)$ is also great, i.e. $P(G_{i}^{(j)} = false|G_{i}^{(j-1)} = true, G_{i-1}^{(j-1)} = false) = r$, where $r > 0.5$. Note that there is still a small chance that gas is still present at $l_i$ at time $t(j)$, which can be a result of a sudden materialization of a new source at $l_i$ or a consequence of local air flow conditions, i.e. $P(G_{i}^{(j)} = true|G_{i}^{(j-1)} = true, G_{i-1}^{(j-1)} = false) = 1 - r$.

The CPT describing the relations between the hypothetical leak and the presence of gas at location $l_1$ at the start of the incident is shown in table II. The CPT reflects the fact, that in case of a leak at location $l_1$ the gas must be present. If the leak is not present, then gas might have been transferred from the neighboring up-wind location $l_E$ relative to the hypothetical leak location. This can happen if gas is present at $l_E$, which is represented by $G_{i}^{(j)} = true$. This is caused by a different source up-wind.

The CPTs describing the dependencies $P(G_{i}^{(j)}|G_{i}^{(j-1)}, L_{(j)}, G_{E}^{(j-1)})$ are not explicitly shown. The CPT has eight columns, where the columns corresponding to $L_{(j)} = false$ are identical to the columns of the CPT $P(G_{i}^{(j)}|G_{i}^{(j-1)}, G_{i-1}^{(j-1)})$. The columns corresponding to $L_{(j)} = true$ are all identical, i.e. $P(G_{i}^{(j)} = true|G_{i}^{(j-1)}, L_{(j)} = true, G_{E}^{(j-1)} = true) = 1$.

The CPTs relating nodes $L_{(j)}$ representing the leak state at location $l_1$ at different time steps are deterministic: $P(L_{(j)} = true|L_{(j-1)} = true) = 1$ and $P(L_{(j)} = false|L_{(j-1)} = false) = 1$. Namely, the model describes a process which would be spawned by a leak lasting for the entire duration of the monitoring process.

Given CPTs shown in tables I and II, respectively, we can easily verify that the causal model in figure 2 captures the notion of a plume where the pollution is irreversible at each
location for a significant period of time. If \( L_i^{(j)} = \text{true} \), i.e. there exists a leak at location \( l_i \) for the duration of the monitoring process, then \( G_i^{(j)} = \text{true} \) for all \( t(j) \). Consequently, after the node \( G_i^{(m)} \) corresponding to the neighboring location \( l_i \) at time \( t(m) > t(j) \) takes on value \( \text{true} \), all subsequent nodes corresponding to \( l_i \) for the time greater than \( t(m) \) take on value \( \text{true} \) as well. The same is true for the following nodes corresponding to other downwind segments. This is a consequence of \( P(G_i^{(j)} = \text{true} \mid G_i^{(j-1)} = \text{false}, G_i^{(j-1)} = \text{true}) = 1 \), which reflects the fact that in case of a plume the gas stays at a certain location until the end of the monitoring process.

The causal model in figure 2 also specifies relations between each \( G_i^{(j)} \) and the observations collected at the corresponding location \( l_i \). Each small network attached to node \( G_i^{(j)} \) captures the causal stochastic process producing observations at location \( l_i \). Note that observation models can be arbitrarily complex causal models themselves, each capturing correlations between different types of observations obtained at a specific location. Little bold circles in figure 2 represent the observations.

B. Simplified Dynamic Models

The aforementioned causal model of the gas propagation and observation generation processes is complex; it is densely connected, which in turn requires expensive, often intractable inference and parameter estimation. A possible solution to this problem proposed in [7] is the use of dynamic BNs, where each time slice is defined over all locations (i.e. segments). This approach, however, can still require large networks if many segments are used. In this paper we introduce an alternative approach to simplifications which explicitly takes into account the characteristics of the modeled process (see section II). The simplified model shown in figure 3 is obtained as follows:

- We introduce nodes \( G_i \), each representing the presence of the gas at location \( l_i \). More precisely, \( G_i = \text{true} \) if the plume front has reached the boundary between segments \( l_i \) and \( l_{i+1} \) and remains so until the end of the estimation process.
- All nodes \( L_i^{(j)} \) and \( G_E^{(j)} \) from the original model in figure 2 are ignored except \( L_i^{(1)} \) and \( G_E^{(0)} \).
- The observation model rooted in a state node \( G_i \) captures all observations that are obtained after the moment at which the plume front has reached the boundary between segments \( l_i \) and \( l_{i+1} \). This can be done because we assume a plume. In other words, the observation models describe how observations are generated in a time interval starting at the passage of the plume front between segments \( l_i \) and \( l_{i+1} \) and the end of the monitoring process.

- Temporal aspects of the gas propagation are not explicitly represented by the model. Instead, temporal aspects are captured through a sequence of segment activations. The evidence at segment \( l_i \) is not processed and fused with the overall model if it is obtained prior to the time at which the plume front would reach \( l_i \) and \( l_{i+1} \). This estimated time can be based on appropriate gas propagation model. After this point in time, all the evidence collected in segment \( l_i \) is treated as if it belonged to a single time slice, which starts at the time of the segment activation. This is a consequence of the fact that in the case of a plume the gas would remain present at each segment until the end of the process. In addition, all evidence collected in segments between the segment containing the triggering report and the hypothesized leak location is immediately used for the instantiation of the observation nodes corresponding to those upwind locations.

- The CPTs \( P(G_i \mid G_{i-1}) \) are simply set to \( P(G_i = \text{true} \mid G_{i-1} = \text{true}) = p \) and \( P(G_i = \text{false} \mid G_{i-1} = \text{false}) = q \), where \( p \) and \( q \) could be taken from the CPT shown in table I.

The resulting simplified model in figure 3 does not capture all causal relations from the original model anymore. However, the simplified model is a good approximation of the distributions over the leaks and the observations, because it captures the most relevant relations represented by the original causal model in figure 2. This can be explained by viewing the observations as outcomes of a forward sampling process on the original model in figure 2. We can show that under realistic circumstances, forward sampling from the complete and the simplified models results in similar constellations of hidden states and observation patterns; i.e. both models capture similar probability distributions over observations and hidden state variables.

We first discuss sampling of the values of hidden state variables \( G_i^{(j)} \) in the original model. We assume that each consecutive state in this model is associated with time \( t(j) = k \cdot t(j)m \), where \( t(j)m \) is the estimated time it would take the plume front to move from the hypothetical leak location \( l_i \) to the boundary between consecutive segments \( l_j \) and \( l_{j+1} \). \( t(j)m \) is computed by using some physical gas propagation model and the assumption that the leak is located at location \( l_i \). Moreover, parameter \( k > 1 \) is a factor which introduces a delay in the segment activation sequence, relative to the timing obtained with the propagation model. Thus, \( t(j) > t(j)m \).

Given a plume originating at \( l_1 \) and the relations between the parameters in the transition CPT from table I, the most likely sequence of state transitions resulting from forward sampling in the original model corresponds to consecutive assignments of value \( \text{true} \) to hidden variables \( G_1^{(1)}, G_2^{(2)}, G_3^{(3)}, \ldots, G_N^{(N)} \); i.e. at each consecutive time
step the variable corresponding to the next segment is instantiated to true. The simplified model explicitly captures only this sequence of transitions between the hidden states.

The model in figure 2 captures also the fact that, occasionally, transitions between consecutive states might be delayed due to various factors that are not explicitly modeled, such as for example obstacles diverting the airflow and slowing down the propagation. In such a case, it might take more than one time step before the variable corresponding to the next segment is instantiated. This is clearly ignored in the simplified model. However, the greater is $k$ the greater is the chance that the state transitions corresponding to instantiation of variables $G_1^{(1)}, G_2^{(2)}, \ldots, G_N^{(N)}$ to true in a sequence will take place. If for a sufficiently great $k$ the most likely transition has not materialized up to a certain point in time, then either the plume does not exist or the source is too weak to cause significant concentrations at the next down-wind location. In other words, with increasing $k$ the distributions over hidden states represented by the original and the simplified models become more similar. Thus, forward sampling on the simplified model would result in similar instantiations of the hidden state nodes as sampling on the original model.

Moreover, given a plume, we see that after a hidden node $G_i^{(j)}$ of the original model was instantiated to true in a sampling process, all nodes $G_i^{(n)}$ with $n > j$ are instantiated to true as well. Thus, after such an instantiation, the sampling of observations at location $l_i$ is always conditioned on $G_i^{(j)} = true$. Since both the original and the simplified models use the same observation models for any location and time step, the conditional distribution from which observations are sampled at specific location are identical for both models after setting $G_i^{(j)} = true$ and $G_i = true$.

Also, contrary to the original model, the simplified model does not capture the materialization of sudden gas sources at locations other than the hypothetical leak location $l_1$. However, this does not have a significant impact on the expected performance, since the prior probability of such events is usually very small.

Finally, the CPT $P(G_i|G_{i-1})$ typically deviates from the true distributions. However, it introduces critical tendencies, such as the fact that the presence of gas at $l_{i-1}$ makes the presence of gas at $l_i$ at the next time step more likely than its absence. Similarly, the absence of gas at $l_{i-1}$ makes the presence of gas at location $l_i$ at the next time step less likely than its absence. Also, the resulting posterior probabilities capture the notion that the reliability of sensor readings for the estimation of a hypothesis is reduced if the readings are located far from the hypothetical source. The robustness with respect to parameters of the CPT $P(G_i|G_{i-1})$ has been confirmed in experiments (see section V-C).

IV. LEAK LOCALIZATION SYSTEM

The leak localization is based on the creation of hypotheses. For each potential source $s$, a hypothesis $h(s)$ assumes that if $s$ were indeed the source, the gas would propagate in a certain way and typical observation sequences (i.e. patterns) in a discretized down-wind area should be obtained. Typical propagation and observation patterns are captured by the simplified domain model presented in the previous sections.

This requires assumption about the downwind area in which the relevant observations for each hypothesis $h(s)$ can be collected. For the sake of simplicity we assume that the gas would spread from the source within an area with a simple form shown in figure 1. Moreover, we assume that the wind has a constant direction and speed. In principle, we could use arbitrary shapes for downwind areas capturing complex dispersion models, such as for example RimPuff [5].

A. Domain Models and Inference

For each hypothesis a specific BN capturing a Hidden Markov process is created. While for each hypothesis we use the same simplified dynamic model capturing the gas propagation, the instantiations based on observations associated with this hypothesis will be specific. Such a BN is used to compute the posterior probability of $P(h|e_h)$ of hypothesis $h$ being true given a set of observations $e_h$ collected in the associated down-wind area $A_h$. In fact, $P(h|e_h) = P(L_t^{(1)}|e_h)$, where $L_t^{(1)}$ corresponds to the location of a hypothesized leak at the estimated time $t(1)$ the plume was released. $t(1)$ is the time of the triggering detection $t(i)$ minus the time the gas would need to travel from the source location $l_1$ to the location of the triggering sensor $l_i$, given the current wind speed. Note that this time computation can be based on much more sophisticated gas propagation models. However, it turns out that for sufficiently high wind speeds, the used model is adequate. The computation of $P(L_t^{(1)}|e_h)$ is based on well known inference algorithms [2]. Moreover, since it is very difficult to estimate the prior probability distributions over the leaks, we simply assume uniform priors. Consequently, the computed posterior $P(L_t^{(1)}|e_h)$ is equivalent to the likelihood of the leak given evidence $e_h$. Note that the influence of priors is reduced as more observations are fused.
B. Data Association

Data association in the presented approach is based on the location of data sources relative to the hypothetical source. In particular, the association is based on a hypothetical plume model originating at the hypothetical source. The plume model provides information on the area in which sensors could be influenced if the hypothetical plume were present. Any measurement obtained within the area in which the minimum detectable concentrations are exceeded is associated with the corresponding hypothesis \( h \) and fused with the help of the dedicated fusion system using the simplified domain model. Note that a single measurement can be associated with multiple hypotheses simultaneously.

However, at an early stage of the detection/localization process we do not know all critical parameters required for a precise estimation of the distribution of concentrations in a plume. Therefore, we use very crude models which just capture the tendencies of the gas propagation processes. We assume that the gas propagates within a cone whose symmetry axis is parallel to the average wind direction measured in the area. The area is subdivided into relatively large segments. In the presented application the distance between the boundaries of a segment \( l_i \) is approximately 1.5 kilometer.

Because of the coarse cone model, information sources which are not influenced by a plume originating in the hypothetical source might be considered relevant and sources that are relevant for a particular hypothesis might be ignored. Thus, the system will be fed by partially misleading observations. Still, by using coarse Bayesian models the impact of occasionally misleading sensor observations can be compensated. In principle, the correct hypothesis will be associated with the greatest probability as long as the cone of the true hypothesis contains the largest number of affirmative observations in the source vicinity. This is also confirmed by experimental results in which the impact of crude plume models was evaluated.

V. Evaluation

In this section we investigate the performance of the presented detection/localization system through a systematic variation of different deviations in controlled experiments. In these experiments we created synthetic sensor data with the help of a simple plume simulation. The synthetic data were fed to a system supporting detection and localization using the simplified model shown in Figure 3.

The performance was defined as the percentage of the estimation points at which the true hypothesis had the highest score, i.e. the highest computed posterior.

A. Setup

Figure 4 presents the used experimental setup: a square of 16 by 16 kilometers containing 287 sensors and five possible sources, numbered 1 to 5 from left to right with number 3 being the true source of a simulated gas plume. The smallest distance between the measurement points is 1 km. Moreover, the wind-speed is assumed constant at 4.5 m/s, and a sampling rate of 3 minutes is used for the measurement devices. The measurement points are represented by 287 squares in a regular grid. Note that a measurement point is present between source 1 and 2 and another between sources 4 and 5 whereas there is no measurement point between sources 2 and 3 and between source 3 and 4. Wind directions are numbered clockwise with 0 being wind blowing toward the bottom of the figure. In order to investigate the impact of geometry, 5 different wind directions are chosen for the evaluation: 270° (wind blowing towards the right hand side of the figure), 250°, 225°, 200° and 180° (wind blowing towards the top of the figure). In general a diffusion cone width of 30 degrees is assumed for the simulation of the gas dispersion. The corresponding cones belonging to the five different wind directions are represented by the full lines, ultra-dash fine, fine dotted, dashed, 3 dash-3 dotted, respectively.

Moreover, each hypothesis is associated with a modeling cone which is divided in 1.5 km wide slices along the wind direction. The angle of the cone is 30 degrees, if not specified otherwise. These cones are used for the data association and the slices correspond to the spatial discretization of the domain models. Due to the space limitations, we discuss three types of experiments. The first type of experiments investigates the impact of sensor noise, the second type focuses on the sensitivity to CPT parameters and in the third set of experiments the impact of the deviations between the coverage of the assumed plume cone and the true plume are discussed.

B. Sensor Noise

In this set of experiments the relations between the sensor noise and the ranking performance were investigated. In Figure 5 the vertical axis represents the average percentage of estimation points at which the true hypothesis had the highest score. The values on the horizontal axis correspond to the sensor noise, i.e. for each sensor the noise corresponds to the chance that this sensor will produce an erroneous reading. Since we only consider sensors with a discrete output, the
evidence injected into the Bayesian fusion system is either confirming or refuting the presence of the gas. Therefore, faulty readings are produced by taking the opposite of the report that would have been correct in a given situation; i.e. when the gas is present then the sensor reports the absence of the gas and vice versa. For each sensor noise level, the detection/localization experiment was repeated 50 times for each of the 5 wind directions. The estimated performance for each direction corresponds to an average over 50 experiments. From figure 5 we can observe that for almost all wind directions (except $270^\circ$) the performance drops when the noise increases. However, wind directions 180°, 200° and 225° show to have very good performance (namely around 80%) even when per sensor the chance of faulty reading is 15%. Also, for 25% noise level for all three wind directions the performance is above 65%. The performance given wind direction 270° is bad. The reason for this is the alignment of the five hypotheses in the direction of the wind. This results in a significant overlap of the five different cones (originating from each hypothesis) which makes discrimination of the true hypothesis from the false hypotheses very hard.

C. Transition Model Deviations

In this set of experiments we investigate the impact of transition models, which are an extremely simple description of the gas propagation mechanism (see section III-B). We used two different CPTs. We first experimented with a CPT defined through $P(G_{i+1} = true|G_i = true) = 0.6$ and $P(G_{i+1} = false|G_i = false) = 0.6$. In the second set of experiments the CPT was defined by $P(G_{i+1} = true|G_i = true) = 0.9$ and $P(G_{i+1} = false|G_i = false) = 0.9$. A rate of noise of 0.15 is used. Moreover, for each CPT the detection/localization experiment was repeated 50 times for each of the 5 wind directions. The estimated performance for each direction corresponds to an average over 50 experiments. Despite significant variations in the used parameters, the performance of the ranking systems using the two different CPTs was not significantly different. For the wind orientations 180°, 200° and 225° there was no observable difference in performance. In the case of the wind direction 250° the performance was around 0.77 in the first experiment and 0.5 in the second experiment. In case of the experiment with orientation $270^\circ$ the performance was around 0.07 in the first experiment and 0.2 in the second experiment. The robustness in the case of directions 180°, 200° and 225° can be explained by the properties of inference with Bayesian networks. Namely, it can be shown that capturing such greater/smaller than relations is sufficient to get a good ranking performance of hypotheses given that for the true underlying model the same relations hold (see paper [8]). Note that, for the ranking no precise probability computation is required.

D. Cone Deviations

In this section we investigate the effect of erroneous cones by increasing or decreasing the cone width (see discussion in section IV-B). Figure 6 shows the sensitivity of the classification performance to different cones. On the horizontal axis we show the error on the cone angle. The cone without an introduced error has a $30^\circ$ angle. For example, an error of $-10^\circ$ means that the cone angle is reduced with $5^\circ$ on either side to $20^\circ$. The vertical axis represents the average ranking performance obtained on the same set of cases all with 15% sensor noise. For each cone deviation, the detection/localization experiment was repeated 50 times for each of the 5 wind directions. The estimated performance for each direction corresponds to an average over 50 experiments.

From figure 6 we see that an error of $-10^\circ$ and wind direction 200° result in good performance, directly followed by a good performance corresponding to wind directions 180°, 225° and 250°. However, when we introduce an error of $+10^\circ$ the performance corresponding to wind direction 200° drops to 20%. Thus, by increasing the cone with $5^\circ$ on either side to $40^\circ$ the ranking performance drops with 60%. The reason for this significant drop in performance is due to the geometry of the sensor grid. The addition of $10^\circ$ results in wider cones corresponding to false hypotheses that include more sensors that affirm the presence of the gas, while the wider cone corresponding to the true hypothesis starts to include sensors that deny the presence of the gas. However, the performance corresponding to wind directions 180°, 225° and 250° remains good for erroneous cones of $-10^\circ$ to $+20^\circ$. Again, we see that wind direction $270^\circ$ results in a bad performance and,
in addition, this performance turns out to be invariant to the considered cone deviations.

VI. DISCUSSION

The paper presents an approach to detection of gaseous substances and localization of the pollution sources using discrete Bayesian networks. Such models can efficiently capture correlations in causal stochastic processes producing heterogeneous, spatially and temporally distributed observations. Contrary to common approaches to source localization, such as for example [3, 4, 10], the presented solution does not rely on concentration/intensity measurements. Namely, in the targeted domains the installation of sufficiently large and dense networks of calibrated chemical sensors are economically infeasible. Therefore, the presented approach supports detection and source localization based on complex patterns of heterogeneous binary observations, such as outputs of different types of chemical detectors based on simple uncalibrated chemical sensors and human reports.

By explicitly taking into account the properties of Bayesian networks and the characteristics of plumes, we derive a simplified model of dynamic gas propagation processes, which supports efficient inference without compromising the overall performance. Since all physical aspects of the plume propagation and the observation mechanisms are usually not known in detail, we use very coarse plume models for (i) the determination of the potential sources, given triggering observations and (ii) the association of data with particular hypotheses. Experimental results show that despite the use of such crude models good overall performance can be achieved.

Moreover, in the targeted domain, it is in general very difficult or even impossible to reliably estimate the conditional probability distributions captured by the CPT parameters in the BNs. However, the used simplified model has a topology for which it was shown that the detection performance can be good and robust against parameter deviations as long as the CPT parameters capture simple greater than smaller than relations [8]. In general, modeling deviations and the lack of modeling granularity are compensated by increasing the numbers of observations. The expected properties of the presented system have been confirmed through experiments, where various deviations were systematically introduced and the impact on the performance was investigated.

The presented gas detection and leak localization techniques are at the core of a recently developed prototype system. The system uses outputs of simple industrial anomaly detectors which process continuous real world sensor data obtained from uncalibrated electronic noses installed in the Port of Rotterdam. In addition, the system can use information provided by humans via interactive voice response system or web pages. In an interactive fashion, the system asks questions about smells and health symptoms that can be answered by yes and no; i.e. we avoid natural language processing. Both, outputs of anomaly detectors and the responses from people are used as evidence in the presented detection and leak localization system. Moreover, the system has been implemented with the help of the Distributed Perception Networks (DPN), a framework which supports distribution of models and inference processes throughout large networks of computing devices [6]. The DPN allows adaptation of the observation models to the changing constellations of information sources [7]. In this way we can efficiently incorporate mobile sensors and adapt the observation models at each location on the fly, as different sources of data become available over time.

With the help of the experiments with synthetic data, we could systematically investigate the interplay between the geometry of sensor networks, various modeling deviations, sensor noise and the performance. These experiments provide a good estimate of the expected performance of the overall system under different conditions, such as sensor locations with respect to the source, specific sensor noise, etc. These experiments are currently being extended to different sensor densities. Also, a preliminary validation of the system is currently being carried out with a few selected sequences of sensor data and human reports collected during real world incidents by the DCMR Milieudienst Rijnmond, an environmental protection agency in the Port of Rotterdam. The used system can process very heterogeneous types of reports. The future research will focus on improved models which can cope with changing winds and puffs.

VII. ACKNOWLEDGMENTS

The presented work is partially funded by the European Union, FP7-grant no. 224318 (DIADEM project).

REFERENCES