Causal Modeling for Fairness in Dynamical Systems (Supplemental)

Elliot Creager [1 2] **David Madras** [1 2] **Toniann Pitassi** [1 2] **Richard Zemel** [1 2]

.

A. Code

Code is available at `github.com/ecreager/causal-dyna-fair` .

B. Experimental Details for Off-Policy Evaluation and Selection

Here, we discuss details on the setup for the off-policy evaluation experiment in Sec. 5.

B.1. Data Generation

We generate data from the Liu et al. (2018) model, described in full in Appendix C. We use $(c_+, c_-) = (75, -150)$ and $(u_+, u_-) = (1, -4)$. We use a single threshold policy of $\tau_j = 620 \forall\ j$. We generate 13 data sets of $10,000$ examples each, using 11 for training (to get confidence intervals), 1 for validation, and 1 for test.

In order to use re-weighting estimators, we must have *overlap* i.e. each point (X, A) must have a non-zero probability of receiving each treatment in the observational data. Since a threshold policy does not satisfy this, we flipped the treatment chosen by the threshold policy with a probability of 0.1.

B.2. Treatment and Outcome Models

We use L2-regularized logistic regression for both the treatment and the outcome model using the `liblinear` default solver in sklearn. We train a treatment and outcome model on each of the 11 training sets, and use these to construct our confidence intervals.

B.3. Estimation of Equal Opportunity Distance

We define the equal opportunity metric δ_{EqOpp} as

$$\delta_{EqOpp} = |P(T = 1|Y = 1, A = 0) - P(T = 1|Y = 1, A = 1)|. \tag{3}$$

The key unit in this expression is $P(T = 1|Y = 1)$ (removing $A = a$ from the right side for clarity). This is non-trivial to estimate, since Y is unobserved for many cases.

We take the following approach. First, using Bayes rule, we have

$$P(T = 1|Y = 1) = \frac{P(Y = 1|T = 1)P(T = 1)}{P(Y = 1)}. \tag{4}$$

$P(T = 1)$ is easy to estimate from observational data. $P(Y = 1|T = 1)$ is the off-policy estimation question — we use either $\mathcal{E}_{Reg}$ or $\mathcal{E}_{DR}$ to estimate this. We estimate $P(Y = 1)$ using off-policy estimation as well, noting that $P(Y = 1) = P(Y = 1|\tilde{T} = 1)$, if $\tilde{T} \perp Y$. Therefore, we can obtain an estimate for the marginal distribution of Y by doing off-policy estimation for random policies $\tilde{T}$ (again, using either $\mathcal{E}_{Reg}$ or $\mathcal{E}_{DR}$). We choose 10 random Bernoulli policies to obtain 10 estimates of $P(Y = 1)$ and average them.

[1]University of Toronto [2]Vector Institute. Correspondence to: Elliot Creager <creager@cs.toronto.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

B.4. Threshold Search

In both the estimation (Fig. 5) and selection (Fig. 6) experiments, we consider all thresholds[1] $\tau \in [300, 850)$ in increments of 5. To choose our best thresholds in the selection experiment, we consider all pairs of group-specific thresholds (τ_0, τ_1), and estimate the value of $\mathcal{V}_\pi$ for the policy associated with those thresholds. We find the optimal value on the validation set, and test them to obtain a final value on the test set, Since we do not require overlap to hold in the target policy, we consider hard threshold policies (we do not flip any predictions post-hoc, as we do in the observational data). In the selection experiment, we test λ in increments of 0.1 from 0 to 0.9.

C. Liu et al. (2018) SCM Details

As briefly discussed above, Liu et al. (2018) propose a one-step feedback model for a decision-making setting then analyze several candidate policies—denoted by the structural equation f_T in our analysis—by simulating one step of dynamics to compute the institution's profit and group outcomes for each policy. Figure 4a shows our SCM formulation of this dynamics model. Here we provide expressions for the specific structural equations used. Throughout, we make the assumption that our model and its associated counterfactuals are representative of the observed data — this is termed as the *consistency* assumption, and is described by Pearl (2010) as

$$P(Y_x = y | Z = z, X = x) = P(Y = y | Z = z, X = x) \tag{5}$$

for all x, y, z, where Y_x is the counterfactual potential outcome for Y under the treatment x.

To sample over $p(X, A)$ we start with Bernoulli sampling of A, parameterized SCM-style like

$$U_{A_i} \sim \text{Bernoulli}(U_{A_i} | \theta); \quad A_i = f_A(U_{A_i}) \triangleq U_{A_i} \tag{6}$$

where $\theta \in [0, 1]$ is the proportion of the $A = 1$ group.

We then sample scores by the inverse CDF trick[2]. Given an inverse cumulative distribution function CDF_j^{-1} for each group $j \in \{0, 1\}$, we can write

$$U_{X_i} \sim \text{Uniform}(U_{X_i} | [0, 1]) \tag{7}$$

$$X_i = f_X(U_{X_i}, A_i) \triangleq \text{CDF}_{A_j}^{-1}(U_{X_i}) \tag{8}$$

Liu et al. (2018) discuss implementing threshold policies for each group $j \in \{0, 1\}$, which are parameterized by thresholds c_j and tie-breaking Bernoulli probabilities γ (for simplicity of exposition we assume the tie-breaking probability is shared across groups). The original expression was

$$\mathbb{P}(T = 1 | X, A = j) = \begin{cases} 1 & X > c_j \\ \gamma & X = c_j \\ 0 & X < c_j. \end{cases} \tag{9}$$

Then, after denoting by $\mathbb{1}(\cdot)$ the indicator function, we can rephrase this distribution in terms of a structural equation governing treatment:

$$U_{T_i} \sim \text{Bernoulli}(U_{T_i} | \gamma) \tag{10}$$

$$T_i = f_T(U_{T_i}, X_i, A_i)$$

$$\triangleq 1^{\mathbb{1}(X_i > c_{A_i})} \cdot U_{T_i}^{\mathbb{1}(X_i = c_{A_i})} \cdot 0^{\mathbb{1}(X_i < c_{A_i})}. \tag{11}$$

A policy f_T (which itself may or may not satisfy some fairness criteria) is evaluated in terms of whether loans were given to creditworthy individuals, and in terms of whether each demographic group successfully repaid any allocated loans on

[1] 300 and 850 are the minimum and maximum credit scores in the dataset

[2] This standard trick is used for sampling from distributions with know densities. Recalling that $\text{CDF}_p : \mathcal{X} \to [0, 1]$ is a monotonic (invertible) function representing $\text{CDF}_p(X') = \int_{-\infty}^{X'} dX p(X < X')$. Then to sample $X' \sim p$ we first sample $U \sim \text{Uniform}(U | [0, 1])$ then compute $X' = \text{CDF}_p^{-1}(U)$.

average. To capture the notion of *creditworthiness*, we introduce a *potential outcome* Y (repayment if the loan were given) for each individual, which is drawn[3] from $p(Y|X, A)$[4]. By convention $T = 1$ as the "positive" treatment (e.g., got loan) and $Y = 1$ as the "positive" outcome (e.g., would have repaid loan if given) Note that Y is independent of T given X, meaning Y is really an indicator of *potential* success. Formally, the potential outcome Y is distributed as $Y_i \sim \text{Bernoulli}(Y_i|\boldsymbol{\rho}(X_i, A_i))$ for some function $\boldsymbol{\rho} : X \times A \to [0, 1]$. We reparameterize this as a structural equation using the Gumbel-max trick[5] (Gumbel & Lieblein, 1954; Maddison et al., 2014):

$$U_{Y_i} \sim \text{Uniform}(U_{Y_i}|[0, 1]) \tag{12}$$

$$Y_i = f_Y(U_Y, X_i, A_i) \triangleq \mathbb{1}\left(\log \frac{\boldsymbol{\rho}(X_i, A_i)}{1 - \boldsymbol{\rho}(X_i, A_i)} + \log \frac{U_Y}{1 - U_Y} > 0\right). \tag{13}$$

The institutional utility u_i and the updated individual score $\tilde{X}_i$ are deterministic functions of the outcome Y_i and the treatment T_i, and the original score X_i:

$$u_i = f_u(Y_i, T_i) \triangleq \begin{cases} u_+^{\mathbb{1}(Y_i)=1} \cdot u_-^{\mathbb{1}(Y_i)=0} & \text{if } T_i = 1 \\ 0 & \text{else} \end{cases}, \tag{14}$$

$$\tilde{X}_i = f_{\tilde{X}}(X_i, Y_i, T_i) \triangleq \begin{cases} X_i + c_+^{\mathbb{1}(Y_i)=1} \cdot c_-^{\mathbb{1}(Y_i)=0} & \text{if } T_i = 1 \\ X_i & \text{else} \end{cases}. \tag{15}$$

As mentioned in Section 5, $\{u_+, u_-, c_+, c_-\}$ are fixed parameters that encode expected gain/loss in utility/score based on payment/default of loan.

There are two *global* quantities of interest. Firstly, the institution cares about its overall utility at the current step (ignoring all aspects of the future), expressed as

$$\mathcal{U} = f_{\mathcal{U}}(u_{1\ldots N}) \triangleq \frac{1}{N} \sum_{i=1}^{N} u_i. \tag{16}$$

Secondly, to understand the societal impact of the lending policy, we measure the average per-group score change induced by the policy, expressed for group $A = j$ as

$$\Delta_j = f_{\Delta_j}(X_{1\ldots N}, \tilde{X}_{1\ldots N}, A_{1\ldots N}) \triangleq \frac{1}{N_{A_j}} \sum_{i=1}^{N} (\tilde{X}_i - X_i)^{\mathbb{1}(A_i=j)}, \tag{17}$$

with $N_{A_j} \triangleq \sum_{i'} \mathbb{1}(A_{i'} = j)$ is the size of the $A_j = 1$ group.

D. Symbol Legends

Here we provide the following symbol decoders for SCMs expressed in the main paper:

- Table 1 decodes the symbols used in Figure 4a
- Table 2 decodes the symbols used in Figure 3

[3]The authors denoted by $\boldsymbol{\rho}(x)$ the probability of potential success at score X. Various quantities were then computed, e.g., $\boldsymbol{u}(x) = u_+\boldsymbol{\rho}(x) + u_-(1 - \boldsymbol{\rho}(x))$. We observe that this is equivalent to marginalizing over potential outcomes $\boldsymbol{u}(x) = \mathbb{E}_{p(Y|X)}[u_+Y + u_-(1 - Y)]$; in our simulations we compute such expectations via Monte Carlo sampling with values of Y explicitly sampled.

[4]The authors use $\boldsymbol{\rho}(X) = p(Y|X)$ in their analysis (suggesting that potential outcome is independent of group membership conditioned on score) but $\boldsymbol{\rho}(X, A) = p(Y|X, A)$ in the code, i.e. the potential outcome depends differently on score for each group. The SCM as expressed in Figure 4a represents the codebase version.

[5]This trick reparameterizes a Categorical or Bernoulli sample as a deterministic transformation of a Uniform sample. See Oberst & Sontag (2019) for discussion of how to perform counterfactual inference for SCMs with Categorical random variables.

Symbol	Meaning
N	Number of individuals
$\|\mathcal{A}\|$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j

Table 1: Symbol legend for Figure 4a

E. Other SCMS

Here we provide some SCMs for some additional papers from the literature:

- Figure 10 describes the multi-step loan setting discussed by Mouzannar et al. (2019). Their model is similar to the one proposed by Liu et al. (2018). The main difference is that Mouzannar et al. (2019) describes dynamics that unfold *exclusively* at the population level, where decisions rendered by the institution do not affect the future well-being of the individuals themselves.

- Figure 11 corresponds to the news recommender simulator discussed by Bountouridis et al. (2019). The goal of this simulator was to understand the long-term effects of recommender algorithms on news consumption behaviors.

- Figure 12 shows the hiring market model proposed in Hu & Chen (2018). Figure 12a shows the higher-level structure of the model: a global state of the hiring market Θ progresses through time, a cohort of workers are initialized at each time step with attributes Φ set by the current global state, and the cohorts progress through time, feeding back into the global state at each step.

  Figure 12b shows the structure of each individual/cohort's journey through the labour market. At the top of Figure 12b, we see the variables which constitute the global state Θ: wages w, reputation Π_μ of group μ, and the proportion of "good" workers on the permanent labour market in group μ, g_μ. The bottom plate of Figure 12b shows the variables which are part of Φ and which correspond to attributes of an individual worker's experience.

Symbol	Meaning
k	indexes groups
P_k	distribution over (X, Y) for group k
b_k	expected group-k baseline population growth at each step
λ_k^t	expected population for group k at time t
α_k^t	mixing coeff for group k at time t
N^t	Total population at time t
Z_k^t	indicator of individual belonging to k-th group
X^t	input features for an individual at time t
Y^t	label for an individual at time t
U_θ^t	Exogenous noise in learning algo. (e.g., random seed)
θ^t	Estimated classifier parameters at time t
$\hat{Y}^t$	Predicted label for an individual at time t
R_k^t	Classification error for group k at time t (unobserved)

Table 2: Symbol legend for Figure 3

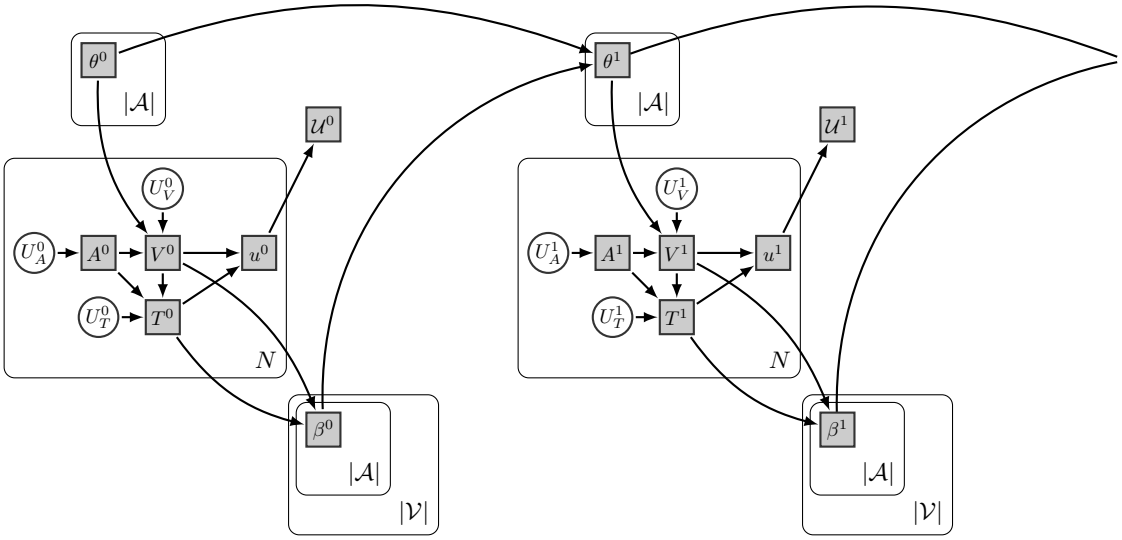


Figure 10: SCM for the group dynamics model proposed by Mouzannar et al. (2019). See Table 4 for a description of each symbol.

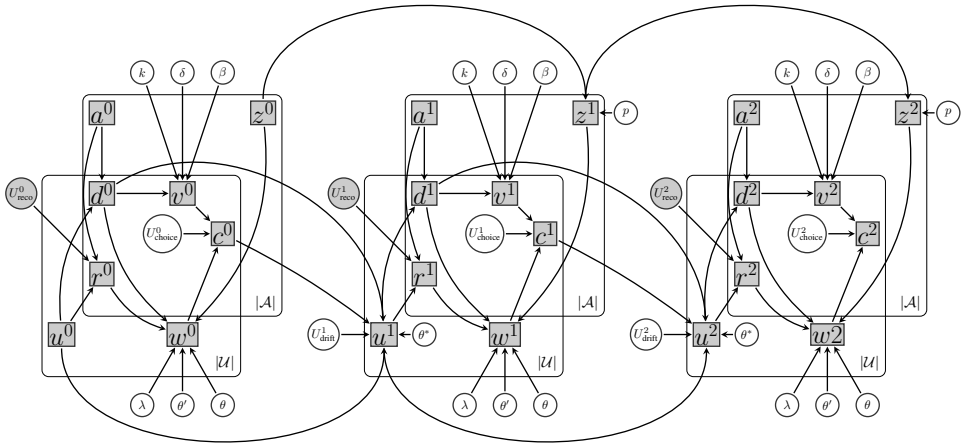


Figure 11: SCM for the news recommendation simulator model proposed by Bountouridis et al. (2019). The key dynamic modeling is in the user vectors in topic space, which drift over time towards the articles that are consumed (these in turn partially depend on the recommendations). Articles are also modeled as decaying in popularity in time. See Table 5 for explanation of all symbols.

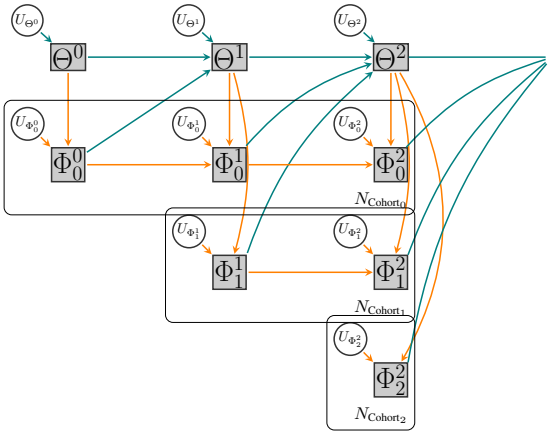
E.1. Symbols for Figures in Supplemental Material

Here we provide the following symbol decoders for SCMs expressed in the Appendices:

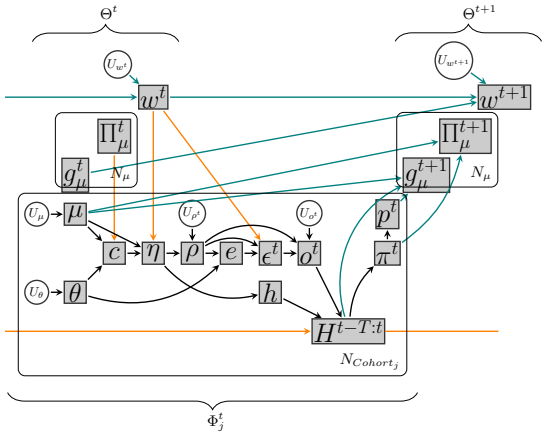
- Table 3 decodes the symbols used in Figure 12
- Table 4 decodes the symbols used in Figure 10
- Table 5 decodes the symbols used in Figure 11

References

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., and Hauff, C. Siren: A simulation framework for understanding the effects of recommender systems in online news environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 150–159. ACM, 2019.



(a) Macro-level DAG showing how market state Θ^t and worker cohorts Φ_j^t dynamically affect one another



(b) Micro-level DAG isolating how market state Θ^t affects investment and effort levels of a single worker cohort, and how worker choices affect market state at the next step

Figure 12: SCM for the hiring model from Hu & Chen (2018). 12a shows macro-level causal assumptions. At step t the global state Θ^t of the PLM affects the choices of all cohorts of workers (a cohort denotes workers that enter the market at the same step) via wage signals (12b). The choices of investment and effort and resulting outcomes in turn affect the workers themselves in terms of hiring decisions, and the global state of the market in terms of average group reputation and performance per group. Teal arrows denote structural functions going into the global state. Orange arrows denote structural functions going into the cohort state. Black arrows denote structural functions within the cohort state. See Table 3 in Appendix D for explanation of all symbols.the dynamics.

Gumbel, E. J. and Lieblein, J. Some applications of extreme-value methods. *The American Statistician*, 8(5):14–17, 1954.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Liu, L., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3156–3164, 2018.

Maddison, C. J., Tarlow, D., and Minka, T. A* sampling. In *Advances in Neural Information Processing Systems*, pp. 3086–3094, 2014.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368. ACM, 2019.

Symbol	Meaning
t	indexes time
i	indexes individuals
j	indexes cohorts
w^t	wages at time t
g_μ^t	proportion "good" group-μ workers in PLM
Π_μ^t	group μ reputation at time t
μ_i	group membership for worker i
θ_i	individual i ability
c_i	cost of investment for individual i
η_i	investment level for individual i
ρ_i	qualification level for individual i
e_i	individual-i cost of effort
ϵ_i^t	individual-i actual effort exerted at time t
o_i^t	individual-i outcome at time t
h_i	was individual hired to TLM following education?
$H_i^{t-\tau:t-1}$	individual-i τ-recent history (outcomes and TLM/PLM status)
π_i^t	individual i reputation at time t
p_i^t	was individual hired to PLM at step t?

Table 3: Symbol legend for Figure 12

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890, 2019.

Pearl, J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*, 21(6): 872–875, 2010.

Symbol	Meaning
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
$\lvert\mathcal{A}\rvert$	Number of demographic groups
V_i	Qualification for individual i
U_{V_i}	Exogenous noise on qualification for individual i
$\lvert\mathcal{V}\rvert$	Number of qualification levels
θ_j^t	Bernoulli parameter of qualifications of group j at time t
N	Number of individuals
T_i	"Treatment" (whether the institution gives loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
$\beta_{j,v}^t$	Selection rate for group j members with qual. v at step t
$\mathcal{U}$	Global institutional utility

Table 4: Symbol legend for Figure 10

	Symbol	Meaning
User	u_i^t	i-th user topic vector at step t
	θ	Awareness decay with user-article distance
	θ'	Awareness decay with article prominence
	λ	Prominent vs proximity in awareness computation
	w	Max awareness pool size for any user
	k	i-th user's sensitivity to article proximity in awareness computation
	θ_i^*	i-th user's sensitivity to article proximity in drift computation
	s	number of articles read per user per step
	$U_{\text{drift},i}^t$	Exogenous noise on user i's drift at step t
	$\lvert\mathcal{U}\rvert$	Number of users
Article	a_j	j-th article topic vector
	z_j^0	initial prominence of article j (possibly shared across topic)
	z_j^t	prominence of article j at step t
	p	prominence (linear) decay factor
	$\lvert\mathcal{A}\rvert$	Number of articles
User-article	$d_{i,j}^t$	distance between user u_i and article a_j at step t
	$v_{i,j}^t$	computed step-t distance (user u_i, article a_j) in awareness computation
	$c_{i,j}^t$	computed step-t choice of user u_i about article a_j
	$U_{\text{choice},i,j}^t$	Exogenous noise on user i's choice of article j at step t
Recommender	m	Number of articles recommended to each user at each step
	$\kappa_{i,j}^t$	Rank of recommendation of article j to user i at step t
	δ	Base amount of salience boost induced by a recommendation
	β	Rank-decay of salience induced by a recommendation
	U_{reco}^t	Exogenous (possibly observed) noise in the recommender algo at step t
	d	Number of steps in the simulation

Table 5: Symbol legend for Figure 11